UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO Y EDUCACIÓN CONTINUA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

HERRAMIENTAS DE APRENDIZAJE DE MÁQUINAS PARA LA PREDICCIÓN DE ACTIVIDADES CRIMINALES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JOAQUÍN ANDRÉS ROA LLANOS

PROFESOR GUÍA:
RICHARD WEBER HAAS
PROFESOR CO-GUÍA:
SEBASTIÁN MALDONADO ALARCÓN

MIEMBROS DE LA COMISIÓN:
JOSÉ PINO URTUBIA
VÍCTOR BUCAREY LÓPEZ

Este trabajo ha sido parcialmente financiado por:
Fondo de Fomento al Desarrollo Científico y Tecnológico

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA Y TESIS PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
Y MAGÍSTER EN CIENCIA DE DATOS
POR: JOAQUÍN ANDRÉS ROA LLANOS
FECHA: 2023
PROF. GUÍA: RICHARD WEBER HAAS
PROF. CO-GUÍA: SEBASTIÁN MALDONADO ALARCÓN

HERRAMIENTAS DE APRENDIZAJE DE MÁQUINAS PARA LA PREDICCIÓN DE ACTIVIDADES CRIMINALES

Al día de hoy, la delincuencia es la principal preocupación de los chilenos. Según el Índice Anual de la Fundación Paz Ciudadana para el año 2022, en uno de cada tres hogares, un integrante de la familia ha sido víctima de hurto, y, según el mismo estudio, el nivel de temor en la ciudadanía ha alcanzado una cifra récord. Frente a la amenaza de la delincuencia, distintas partes involucradas han comprometido avances en su combate, y la necesidad de innovación se hace imperante. En el contexto de la seguridad ciudadana y vecinal, SOSAFE surge como una alternativa para la prevención del crimen. En la aplicación, usuarios emiten reportes de todo tipo, principalmente vinculados a la seguridad y actividades sospechosas observadas en el barrio.

Con el fin de contribuir a la prevención de actividades de carácter criminal, para la presente investigación de tesis se implementaron herramientas de aprendizaje de máquinas para la predicción de actividades criminales en la Región Metropolitana de Santiago, utilizando reportes de SOSAFE. Nunca antes se ha publicado un estudio en el cual reportes directos de ciudadanos, vinculados a actividades sospechosas observadas en el barrio, contribuyan a la elaboración de un modelo de predicción del crimen. Para la elaboración de los modelos predictivos, se dispuso de reportes categorizados de la aplicación, asociados a una determinada latitud, longitud, fecha y hora, entre otros, para un período total de dos años.

Se agruparon los reportes, según las distintas categorías a las que pertenecían, bajo distintas configuraciones espaciales y temporales. El objetivo fue predecir la categoría asociada al crimen, para un período siguiente, en las grillas definidas a nivel espacial, utilizando como principal atributo para la predicción la cantidad de reportes pertenecientes a todas las otras categorías en períodos previos, además de variables socioeconómicas complementarias.

Se utilizaron seis métodos de clasificación binaria. Los mejores resultados se obtuvieron para el modelo *XGBoost*, para la configuración de submuestreo de los datos, para una configuración de ventanas temporales de dos semanas, y grillas espaciales de aproximadamente un kilómetro cuadrado de área. Dicho modelo reportó una *accuracy* de $0,8223$, un F–*Measure* de $0,8348$ y un puntaje ROC–AUC de $0,8365$.

Se concluyó que el modelamiento propuesto tiene un buen desempeño y escalabilidad, prediciendo con una precisión adecuada los reportes asociados a actividades criminales en la aplicación, identificando la importancia de los atributos más relevantes para la predicción, de esta manera contribuyendo a la prevención del delito en un ámbito táctico y estratégico, al posibilitar la localización de recursos para el combate a la delincuencia, facilitando el acceso al inmediato sentir ciudadano, y sentando la base de futuros trabajos en esta dirección.

A mi familia

Agradecimientos

Quiero agradecer a mi mamá, a mi papá y a mi hermano, por su apoyo incondicional, y en todo momento. A mi tía Ana por su recibimiento y su cariño. A mis abuelos, mis tías y tíos, y a mis primos, por recibirme, incorporarme, y hacerme sentir bienvenido en la ciudad.

Agradezco profundamente a mis amigos y compañeros de la Sección 2 y Plan Común, en particular a Nicolás, Diego, Maximiliano, Mauricio, Mariana, Felipe, y a tantos otros que me acompañaron durante tantos años. También a mis amigos de especialidad, a Patricio, Pablo y Francisco, y a todos quienes me apoyaron en el rigor universitario, y en tantas tardes prolongadas de estudio. Agradezco también a Fernanda por su amistad, a Matías por ayudarme en la aventura del magíster, y a Fernando, por su apoyo como padrino.

Mis agradecimientos al profesor Richard Weber por aceptarme en su equipo de trabajo, y por su voluntad para guiar mi tesis. Al profesor Sebastián Maldonado, por su colaboración y excelente disposición para co-tutelar mi investigación. A los profesores José Pino y Víctor Bucarey por aceptar ser parte de la comisión. Agradezco a todo el equipo FONDEF por integrarme en este proyecto, en particular, mi agradecimiento a Florencia, Carla, Matías y Yerko. Gracias a Armadillo y a Sebastián por su ayuda en la redacción de este documento.

También quisiera agradecer a quienes hicieron posible mi pasantía en el extranjero bajo el marco de este proyecto de investigación. Al profesor Kristof Coussement, mi más grande agradecimiento, por aceptar guiarme bajo su tutela. Agradezco a mis compañeros de la oficina de postgrado, y a mis amigos y compañeros de piso, por acompañarme y regalarme experiencias que atesoro. Quiero agradecer también a la Vicerrectoría de Asuntos Académicos de la Universidad de Chile y al Instituto Francés por otorgarme la beca de investigación.

Mi agradecimiento a toda la comunidad universitaria de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, del departamento de Ingeniería Industrial, y a la coordinación del magíster. Quisiera agradecer a los profesores que durante mi estancia me aceptaron como miembro de sus equipos docentes, en particular, a los profesores Aidan Hogan, Andrés Fernández y Christian Diez, por, además, recomendarme para el programa de magíster, y para mi pasantía en el extranjero. Agradezco también a la profesora Constanza Contreras el haberme otorgado la oportunidad de tener mi primera experiencia docente.

Por último, envío finalmente mi agradecimiento a quienes trabajan informalmente todos los días a mediodía a las afueras del campus, quienes, pese a trabajar en condiciones tremendamente inestables, se han preocupado a lo largo de mi estancia en la universidad, y en numerosas oportunidades, de entregarme un sincero gesto de apoyo cuando el ritmo académico se ha tornado difícil. Desde este último espacio, mi humilde reconocimiento.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

Durante el último tiempo, la delincuencia, la criminalidad y la victimización han sido temas de alta repercusión e interés nacional.

A fines de octubre de 2022, la Fundación Paz Ciudadana presentó los resultados de su Índice 2022 [1], indicando que los ciudadanos que se clasifican en un nivel de temor alto alcanzaron una cifra récord en la historia de dicha medición, llegando a un 28 % a nivel nacional, donde las mujeres se declaran como las más afectadas (35,4 %). El temor medio, según dicha medición, alcanzó un 70,4 % de los encuestados. La fundación declaró también que un 59 % de los encuestados ha reforzado la seguridad de su casa, un 71 % ha dejado de salir a ciertas horas y un 75 % ha dejado de salir a ciertos lugares por miedo a ser víctima de la delincuencia. En esa misma línea, la Encuesta Nacional Urbana de Seguridad Ciudadana, aplicada anualmente por el Instituto Nacional de Estadísticas, informaba en agosto de 2022 que la tasa de percepción de aumento de la delincuencia en el barrio donde viven los encuestados llegaba al 45,8 % [2].

En cuanto al delito de robo, el Índice 2022 de la Fundación Paz Ciudadana cuantificaba la victimización en un 31,6 %, lo que significa que aproximadamente en uno de cada tres hogares un integrante de la familia ha sido víctima de hurto. Además, según la información publicada por el Sistema Táctico de Operación Policial de Carabineros de Chile, el robo de vehículos creció en un 42 % durante el primer semestre de 2022, en comparación con el primer semestre de 2019. En particular, el robo violento de vehículos creció en un 81 % [2].

El aumento de los índices de desempleo y de pobreza, situación observada luego de la última pandemia sanitaria, posee una notable correlación con el aumento de los índices de delincuencia [3]. Se espera que la delincuencia aumente hacia el 2023, en conjunto con la posible agudización de una potencial crisis económica [4].

Todos estos antecedentes, junto a distintos factores de la agenda nacional, han colocado a la delincuencia como tema fundamental para la ciudadanía, las autoridades y para las distintas partes interesadas.

En respuesta a las mediciones publicadas en octubre de 2022, el Presidente de la República, Gabriel Boric Font, declaró "que personas dejen de hacer cosas porque tienen miedo, nos obliga como Estado a realizar una tarea con mucha más fuerza. Tenemos que recuperar nuestros espacios públicos de la delincuencia y se tiene que hacer una labor conjunta liderada por el Estado, fortaleciendo a las policías, pero también en conjunto con el sector privado que, en esto no me cabe ninguna duda, son aliados". En la misma línea, el alcalde de Independencia, Gonzalo Durán Baronti, propuso "incrementar la presencia de los órganos del Estado, incluidas las policías, con el propósito de contribuir a bajar esa sensación de temor generalizado" [5].

En ese sentido, luego de que la primera Ministra de Interior y Seguridad Pública, Izkia Siches Pastén, declarara en mayo de 2022 que "las policías y la seguridad son un desafío de Estado", el Presidente de la República anunció un Plan Nacional de Seguridad que contempla, entre otras medidas, una estrategia de recolocación e incorporación de agentes de Carabineros en los sectores más críticos del país [6].

Distintas voces autorizadas han planteado la necesidad del Estado de hacerse cargo de la situación actual de la delincuencia en Chile, y de innovar en posibles soluciones. El académico Mauro Basaure planteaba en octubre de 2022 que "cuando las políticas públicas no están teniendo los resultados esperados, y se enfrentan escenarios de crisis de seguridad y de potencial peligro antidemocrático, la innovación se vuelve obligatoria". En sus palabras, "la situación del país torna urgente la innovación en los modos de atacarla" [4].

1.1.1. Aplicación móvil y origen del conjunto de datos

SOSAFE es una aplicación móvil orientada a la seguridad vecinal. En dicha aplicación cada usuario puede emitir un reporte, de manera individual, al cual luego podrán acceder los vecinos de su barrio, ciudadanos en general, y autoridades pertinentes.

El contexto de los reportes emitidos es variado. La aplicación fue creada pensando en los reportes en materias de seguridad, pero al día de hoy pueden reportarse desde actividades sospechosas, hasta venta de repostería a nivel vecinal, pasando por reportes de mascotas perdidas, delincuencia en general, y temáticas varias asociadas a la vida en comunidad.

En el contexto de la seguridad ciudadana, SOSAFE puede resultar de gran ayuda para las autoridades en materia de seguridad, tanto a nivel local de municipalidades, como a nivel centralizado, como Carabineros, Policía de Investigaciones y las distintas Fiscalías a nivel nacional. Los reportes ciudadanos dan cuenta en numerosas oportunidades de crímenes cometidos o por realizar, indicando antecedentes claves que luego pueden ser recogidos por las autoridades para el desarrollo de sus funciones en materia de prevención y seguridad.

La experiencia de los usuarios, en el contexto de seguridad, es la siguiente. Un usuario de la aplicación móvil, al presenciar una actividad sospechosa de cualquier tipo, o habiendo atestiguado la concreción de un crimen, escribe en la aplicación móvil un reporte breve de lo que ha visto, explayándose a su parecer, e incorporando la información que le parezca relevante. Asigna el reporte manualmente a una categoría predeterminada, y le asigna también un sector de la ciudad como referencia. Finalmente, envía el reporte. La hora, fecha y ubicación exacta del reporte quedan también registradas.

1.1.2. Acercamiento al problema mediante ciencia de datos

La ciencia de datos – o *data science*, en inglés – es una disciplina que se relaciona a conceptos tales como la inteligencia artificial y el aprendizaje de máquinas, y está fuertemente vinculada a la teoría matemática, a la computación, a la programación y al diseño de algoritmos. Modelos basados en ciencia de datos pueden resultar en un inmenso beneficio para una numerosa cantidad de rubros, dado el valor científico–matemático que puede incorporar la ideación de este tipo de soluciones. Dentro de lo que es la ciencia de datos, el aprendizaje de máquinas – o *machine learning* en inglés – es el campo de estudio que le otorga a las máquinas la habilidad de aprender sin ser explícitamente programadas [7]. Mediante la incorporación adecuada de conjuntos de datos, y el aprendizaje de patrones presentes en los mismos, se puede llegar a lo que se denomina 'inteligencia artificial', es decir, que las máquinas desarrollen la capacidad de tomar decisiones que en otro contexto podrían considerarse decisiones tomadas por un humano.

Modelos basados en ciencia de datos permiten obtener conclusiones, las cuales pueden utilizarse para la implementación de decisiones de gestión a corto, mediano o largo plazo. Para las organizaciones puede resultar de mucho valor poder sustentar sus decisiones, operacionales o estratégicas, en conclusiones obtenidas mediante modelos de valor científico–matemático. Este paradigma tiene validez tanto para el sector privado – donde organizaciones del rubro bancario, del *retail* o telecomunicaciones, entre muchas otras, invierten grandes cantidades de dinero en el desarrollo de modelos predictivos basados en aprendizaje de máquinas – como para el sector público, donde miles de decisiones de políticas públicas pueden verse beneficiadas por la implementación de modelos de ciencia de datos.

No son pocas las oportunidades en que partes interesadas en la reducción del crimen han desarrollado modelos computacionales para dicho fin. En todo el mundo, policías, gobiernos y fiscalías han hecho avances no sólo en sistemas y modelos computacionales orientados a la seguridad, sino que en la mecanización de sus procesos en general. Han incorporado valor computacional, matemático y científico en el desarrollo de sus funciones, optimizando su carga laboral y mejorando la escalabilidad de dichas funciones. Para las comunidades implicadas en la toma de decisiones de políticas públicas por parte de estas instituciones también puede resultar conveniente una correcta implementación de dichos programas, ya que puede resultar en una mejora directa de su calidad de vida y su percepción de seguridad.

Investigaciones previas no permitieron dar cuenta de ningún estudio existente que vincule reportes criminales de carácter ciudadano, en aplicaciones móviles de las características descritas para SOSAFE, con modelos de predicción del crimen, basados en reportes de actividades sospechosas en el barrio o de otros tipos semejantes. Los modelos de predicción del crimen han mostrado un incremento numérico considerable durante los últimos diez años, acompañados de los avances computacionales y en inteligencia artificial, por lo que la contribución del presente estudio aborda una arista aún inexplorada en este joven campo de estudio. Usualmente, los estudios cuantitativos para la predicción del crimen se basan en reportes policiales, o en otros datos afines, pero escapan de la percepción ciudadana inmediata, aspecto que los reportes de SOSAFE capturan en una plataforma especialmente diseñada para dichos fines.

1.2. Objetivos

1.2.1. Objetivo General

Construir un modelo predictivo de clasificación binaria, que permita predecir si ocurrirán o no actividades de carácter criminal en un cierto lugar y período de tiempo, en la Región Metropolitana, en base a reportes de SOSAFE.

1.2.2. Objetivos Específicos

- Proponer una o más configuraciones a nivel espacial de los datos, que permitan agrupar los reportes en la Región Metropolitana de manera tal de implementar modelos de clasificación binaria para obtener conclusiones respecto al problema de investigación
- Agrupar los reportes disponibles, distinguiendo por cada categoría, a nivel temporal, proponiendo una o más configuraciones considerando la pertinencia para el caso y la naturaleza de los datos.
- Construir un conjunto de datos en base a los datos proporcionados por la empresa, otras fuentes de datos complementarias y las configuraciones espaciales y temporales ideadas, que permita implementar modelos predictivos de clasificación binaria y obtener resultados pertinentes al problema de investigación.
- Implementar algoritmos predictivos de clasificación binaria utilizando el conjunto de datos construido para el caso.
- Implementar métricas de evaluación que permitan concluir sobre el desempeño de los modelos generados

1.2.3. Alcances

Las herramientas de aprendizaje de máquinas son muchas y muy variadas, y pueden ocuparse para fines muy distintos. En una subsección posterior se mencionarán las herramientas que existen, y se describirán las que para este proyecto se utilizan, pero por ahora cabe destacar que son algoritmos binarios de clasificación, dejando de lado otras herramientas tales como análisis de regresión y modelos de aprendizaje no supervisado. En ese mismo sentido, se privilegiará el uso de algoritmos clásicos de clasificación binaria que no requieran el uso intensivo y extenso de grandes conjuntos de datos ni de recursos computacionales, por lo que no se profundizará en el desarrollo de redes neuronales profundas. Con el fin de profundizar el desarrollo de los mencionados modelos de clasificación binaria, no se utilizan en este trabajo herramientas de análisis de texto, ni de procesamiento de lenguaje natural, escapando de los alcances del presente trabajo investigativo.

Se entiende, además, por análisis de actividades criminales a aquellas que se encuentran contenidas dentro de las categorías pertinentes, en los reportes disponibles de SOSAFE. Delitos de carácter económico, de violencia de género, u otros, no entran en el alcance de la presente investigación. Las categorías asociadas a actividades delictivas dentro de la aplicación se detallan en la sección 3.1.1.

Capítulo 2

Marco Teórico

En el siguiente capítulo se presentan los conceptos más relevantes para la comprensión del desarrollo del caso de estudio.

En primer lugar, se introducen y detallan los conceptos vinculados al aprendizaje de máquinas e inteligencia artificial, con un mayor foco en los aspectos relevantes para el objetivo general, como lo son los modelos de clasificación binaria. Se presentan también las principales métricas de desempeño para la evaluación de dichos modelos, con principal foco en la aplicación al presente caso.

El capítulo continúa con una revisión bibliográfica. Se presentan los principales trabajos relevantes en materia de predicción del crimen, y se concluye sobre los principales hallazgos y metodologías, dependiendo de la naturaleza de los datos y del tipo de problema a resolver. Se presenta también el estado del arte y la vanguardia en la materia.

La segunda parte del capítulo introduce conceptos relevantes respecto al origen del conjunto de datos. En primer lugar, se incluye una caracterización de la empresa, con foco en su modelo de negocios y posicionamiento estratégico. Se muestran ejemplos de reportes de la aplicación móvil y se introduce el conjunto de datos disponible. Luego, se presenta el espacio geográfico relevante para el caso, el cual corresponde al territorio nacional y sus divisiones político-administrativas. Finalmente, y en vista de los antecedentes presentados, se concluye sobre los sesgos naturalmente presentes en el conjunto de datos y que podrían resultar relevantes para la elaboración de los modelos.

2.1. Aprendizaje de Máquinas

2.1.1. Inteligencia Artificial

2.1.1.1. Contexto

La inteligencia artificial es una disciplina enigmática [8]. Es una disciplina de casi setenta años de edad cuya definición aún no converge a estándares comúnmente aceptados. Para definir inteligencia artificial es necesario embarcarse en la complejísima tarea de definir 'inteligencia' y, además, definir 'artificial'.

A grandes rasgos, se habla de que la inteligencia artificial o computacional, al contrario de la inteligencia natural, es el campo de estudio del diseño de agentes inteligentes [9]. Es decir que, al contrario de aquellos seres que poseen cerebros complejos y orgánicos, como los animales - y por cierto, los humanos - la inteligencia artificial hace referencia al estudio de componentes artificiales que poseen la capacidad de 'razonar'.

La inteligencia artificial, por tanto, resulta ser una disciplina sumamente amplia, comprendida por campos de estudio muy variados como lo son la psicología, la filosofía, la biología y la historia, y, para los fines que este texto refiere, también la matemática, la computación y las políticas públicas.

2.1.1.2. Origen

Históricamente, el origen de la inteligencia artificial como disciplina se remonta a la década de los cincuenta. Durante esta década, se llevaron a cabo muchos avances a niveles algorítmicos y computacionales, de manera simultánea.

Alan Turing en 1950 formuló la pregunta "¿Pueden pensar las máquinas?" [10]. Para responderla, construyó una prueba que hoy es llamada el 'Test de Turing', en el cual un interrogador busca descifrar la identidad y naturaleza de su contraparte mediante una serie de preguntas. Mediante su prueba se definió por muchos años a la disciplina de la Inteligencia Artificial como aquella en la cual una máquina buscaría hacerse pasar por un humano, engañando a un interrogador imparcial al tomar una serie de decisiones que parecerían de naturaleza humana.

Más adelante, y de manera independiente, fue en 1956 cuando John McCarthy inventó el término de "inteligencia artificial", en la conferencia de Dartmouth, al definir los trabajos que múltiples profesionales del rubro de la computación habían estado llevando a cabo durante los cinco años previos.

En 1959, Arthur Samuel tuvo también uno de los acercamientos más tempraneros y relevantes al estado del arte, escribiendo un programa de computación con la capacidad de aprender a jugar damas [7]. Samuel diseñó un algoritmo que le permitía a su programa calcular los costos y beneficios de las jugadas posteriores al movimiento de un rival, por lo que, a diferencia de, por ejemplo, una calculadora moderna, el programa era capaz de proponer jugadas y respuestas que nunca antes un humano le había explícitamente indicado que realizara. El programa de Samuel fue precursor de los innumerables programas que, pocos

años después, campeonaron en ajedrez derrotando a profesionales jugadores humanos. Fue precursor de las múltiples aplicaciones de inteligencia artificial que al día de hoy ejercen, por ejemplo, recomendaciones de manera automática en los teléfonos celulares, y en los sistemas automatizados de las grandes industrias.

Para los fines del presente texto, el foco en la investigación de Samuel es la manera en la cual le enseñó a su programa a alcanzar la 'inteligencia artificial'. La disciplina que le enseña a las máquinas a obtener la inteligencia artificial se llama "aprendizaje de máquinas".

2.1.2. Aprendizaje de Máquinas

Mientras que la inteligencia artificial hace referencia a la amplia ciencia de la imitación de las habilidades humanas, el aprendizaje de máquinas (o aprendizaje automático) es un subconjunto específico de la disciplina, que se enfoca en el entrenamiento de las máquinas para su aprendizaje [11].

Tradicionalmente, el aprendizaje de máquinas se ha dividido en dos subáreas; el aprendizaje supervisado y el aprendizaje no supervisado. El aprendizaje automático supervisado consiste en modelos matemáticos y computacionales que permiten predecir una etiqueta o valor en base a observaciones previas. Por su parte, el aprendizaje automático no supervisado no presenta valores o etiquetas a predecir, y su objetivo es encontrar patrones en los datos. Ejemplos de modelos de aprendizaje automático no supervisado son los modelos de *clustering*, o los modelos para la reducción de la dimensionalidad, tal como el análisis de componentes principales.

Generalmente, se identifican dos tipos de modelos de aprendizaje automático supervisado, los modelos de regresión y los modelos de clasificación.

2.1.3. Modelos de clasificación

Los modelos de clasificación son modelos predictivos de ciencia de datos en los cuales se busca clasificar una observación según los datos que se tienen de ella. A diferencia de modelos de regresión, que buscan predecir cantidades numéricas continuas, los modelos de clasificación buscan predecir clases, categorías o etiquetas.

Pueden clasificarse las observaciones de un cierto conjunto de datos de manera binaria, o bien pueden clasificarse observaciones dentro de un conjunto de categorías, lo que se conoce como clasificación multiclase. Las categorías no necesariamente tienen un orden definido, pero bien podrían tenerlo.

Se listan a continuación algunos clasificadores binarios, sin discriminar orden de importancia.

1. **Regresión Logística** (también conocida como *Logistic Regression* $- LR$)
2. **Máquinas de vectores de soporte** (*Support Vector Machines* $- SVM$)
3. **Naïve Bayes** (NB)
4. $k-$**vecinos más cercanos** (KNN)
5. **Árboles de Decisión** (*Decision Trees* $- DT$)
6. **Random Forest** (RF)
7. **XGBoost** (XGB)
8. **Perceptrón multicapa** (*Multi–Layer Perceptron* $- MLP$)

Los clasificadores anteriormente listados son algoritmos diseñados para la clasificación binaria. También son matemáticamente extensibles para el caso multiclase, ya sea mediante heurísticas, o de manera intrínseca, dependiendo del clasificador. La clasificación binaria, por tanto, se realiza aplicando dichos algoritmos, de manera individual o paralela, a los conjuntos de datos especialmente acomodados para ello.

2.1.4. Redes neuronales artificiales

A mediados del siglo pasado ocurrió un aumento en el desarrollo de modelos computacionales basados en la biología y en la naturaleza. Esto se observó, por ejemplo, en el desarrollo de los algoritmos genéticos, popularizados por John Henry Holland [12], y en la invención de uno de los modelos de clasificación binaria más importantes de la historia; el perceptrón. Éste algoritmo fue introducido por Rosenblatt en 1958, basado en la composición de las redes neuronales [13]. Un perceptrón busca ser un símil matemático y artificial de una neurona. Una composición adecuada de perceptrones forma una red neuronal artificial, que a su vez funciona como un clasificador binario, extensible al caso multiclase. Las redes neuronales artificiales son el fundamento de una disciplina conocida como aprendizaje profundo o *deep learning*.

Las redes neuronales artificiales han sido ampliamente utilizadas en la literatura de modelos predictivos, sobre todo durante los últimos años, en primer lugar, por una serie de características matemáticas ventajosas de los modelos, y, por otra parte, por los avances computacionales de los últimos años que han vuelto factible su implementación. Su uso e implementación se asocia a modelos de un gran desempeño y versatilidad, pero asimismo, requieren de un cuantioso volumen de datos disponibles para el entrenamiento de estos modelos. Esto también se puede observar en el caso particular de la literatura de predicción del crimen, donde numerosos modelos y distintas arquitecturas han sido propuestas con fines de la predicción y prevención del crimen.

2.1.5. Métricas de desempeño

Los resultados de un modelo de ciencias de datos son sometidos a evaluación mediante métricas especialmente construidas para medir su desempeño. En el caso de los modelos de clasificación, estas métricas suelen ser *Accuracy*, *Recall*, *Precision* y F1–*Measure* (o F1–*Score*). Todas estas métricas derivan de, o se relacionan a, la matriz de confusión, la cual es un diagrama que permite verificar la cantidad de predicciones correctas que realiza un modelo.

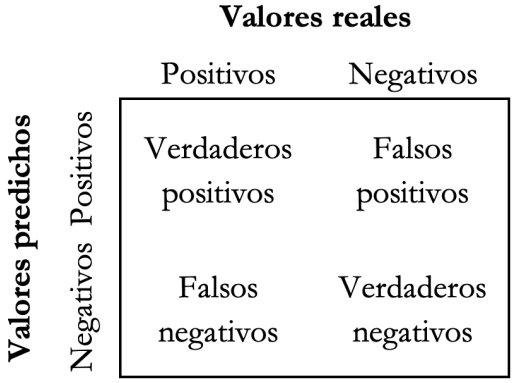

Figura 2.1: Matriz de confusión

En la Figura 2.2 puede observarse un diagrama de una matriz de confusión para un caso binario. Se define, arbitrariamente y según el caso de uso, a un valor a predecir como la clase 'positiva' y a la contraria como la clase 'negativa'. La matriz tiene cuatro regiones, catalogadas como cada uno de los cuatro casos que permite la predicción de un modelo binario; los 'Verdaderos Positivos', 'Verdaderos Negativos' para las predicciones correctas de cada una de las dos etiquetas posibles, y 'Falsos Negativos' y 'Falsos Positivos' para las predicciones erróneas del modelo. Estas cuatro regiones de la matriz se llenan cada una de la cantidad de predicciones para cada uno de los casos.

En el caso particular de los modelos de predicción del crimen, existen altos incentivos para mejorar el desempeño de los modelos. Predicciones erróneas pueden resultar en la asignación incorrecta de recursos públicos o policiales, o la toma de decisiones públicas incorrectas y que perjudiquen a las comunidades.

No es necesario elegir sólo una métrica de desempeño, así como tampoco es necesario usar una gran cantidad. Las métricas a utilizar dependen del foco del problema y del caso de uso en particular.

Accuracy

La *Accuracy* es la métrica de desempeño que calcula la razón entre las predicciones correctas (tanto de la clase positiva como de la negativa) y los casos totales [14].

La fórmula es descrita por la Ecuación 2.1, donde VP se refiere a la cantidad de 'Verdaderos Positivos', VN a los 'Verdaderos Negativos', FN a los 'Falsos Negativos' y FP a los 'Falsos Positivos'.

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.1}$$

Precision

En ocasiones, y dependiendo del caso de uso, puede resultar tanto o más relevante no sólo evaluar los modelos según la proporción de predicciones correctas para ambas clases dividido por el número de casos totales, sino que también puede resultar pertinente evaluar los modelos enfocándose en una subsección en específico de la matriz de confusión. Así, la *Precision* evalúa los modelos según la razón entre los casos determinados como verdaderos positivos dividido entre el total de predicciones para la clase positiva, tanto correctas como incorrectas. La clase determinada como positiva es arbitraria, y está dada según el problema a resolver.

La fórmula para calcular la *Precision* está dada por la Ecuación 2.2. En esta ecuación, así como en las siguientes que se presentan en esta subsección, se presentan las mismas abreviaciones introducidas en la Ecuación 2.1.

$$Precision = \frac{VP}{VP + FP} \tag{2.2}$$

Recall

El cálculo del *Recall* es análogo al de *Precision*, pero a diferencia de aquella métrica, en vez de evaluar los modelos según las observaciones predichas como positivas, se enfoca en las observaciones originalmente positivas, ponderando las correctamente categorizadas con respecto a aquellas observaciones predichas por el modelo a evaluar como pertenecientes a la clase positiva o negativa. Es importante notar que tanto la medida de *Recall* como la de *Precision* fueron introducidas de manera complementaria, en el contexto del campo de la recuperación de la información, por Gerald Salton [15].

La fórmula para calcular el *Recall* está dada por la Ecuación 2.3.

$$Recall = \frac{VP}{VP + FN} \tag{2.3}$$

F1–*Measure*

El F1–*Measure* o F1–*Score* es una métrica que combina y pondera *Precision* y *Recall*. La fórmula tradicional está dada por la Ecuación 2.4, sin embargo, puede variar según ponderaciones arbitrarias que se les quiera asignar a cada una de las dos métricas mencionadas.

$$F1 - Measure = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{VP}{VP + 0,5 * (FP + FN)} \tag{2.4}$$

Curva ROC y AUC

La curva ROC (*Receiver Operating Characteristic*) es un diagrama que compara la tasa de Verdaderos Positivos (*Recall*), en el eje de las ordenadas, contra la tasa de Falsos Positivos, en el de las abscisas.

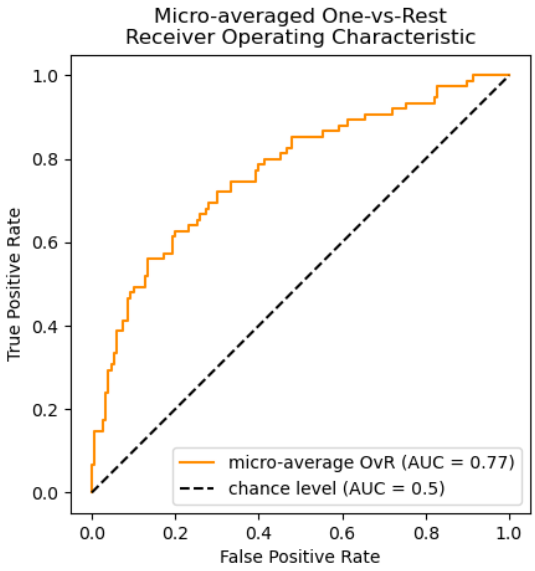

Figura 2.2: Ejemplo de implementación de diagrama ROC en la librería *scikit-learn* de Python [16]. Un buen modelo predictivo tenderá a trazar una curva que estará cerca de la esquina superior izquierda.

Desde la curva ROC surge la métrica AUC (*Area Under the Curve*), la cual, al igual que las cuatro métricas de desempeño anteriormente mencionadas, es un valor que va desde 0 a 1, donde valores más cercanos a 1 indican mayor cercanía a un 100 % de aciertos para las predicciones de la clase positiva, y 0 % de error en la clase negativa. Ésta, tal como el nombre en inglés lo indica, puede medirse calculando el área bajo la curva ROC.

2.1.6. Balanceo del conjunto de datos

Existe la posibilidad de que los conjuntos de datos con los cuales se desarrollan modelos predictivos de ciencias de datos se encuentren desbalanceados. En un caso de clasificación binaria, esto quiere decir que una clase a predecir tiene más representatividad en el conjunto de datos que la otra. Cuando este nivel de representatividad excede un cierto umbral − por definir −, y queda en evidencia que una clase está sobrerrepresentada, los modelos de clasificación pueden verse perjudicados en su desempeño.

Para solucionar este desperfecto, es posible implementar técnicas de balanceo de datos. Principalmente son dos; *oversampling*, o sobremuestreo, y *undersampling*, o submuestreo. Mientras que la primera técnica crea artificialmente observaciones de la clase subrepresentada para equipararla a la contraria, la segunda elimina observaciones de la clase sobrerrepresentada, para que así ambas cantidades queden balanceadas.

Una de las implementaciones más populares de *oversampling* se llama *SMOTE* [17]. A diferencia del *oversampling* tradicional, que agrega observaciones preexistentes en el conjunto de entrenamiento al azar, *SMOTE* es capaz también de 'inventar' observaciones nunca vistas, pero que según su algoritmo hagan sentido, tal como podría ser una medición continua de peso o de estatura, previamente inexistente, entremedio de dos observaciones disponibles en el conjunto de datos.

2.2. Predicción del crimen

El estudio del comportamiento criminal, desde una perspectiva de la prevención y predicción del delito, ha sido abordado desde múltiples rubros y campos de estudio, contando desde la sociología, la psicología y la geografía, hasta las matemáticas y las ciencias de la computación, entre muchos otros.

2.2.1. *Hotspots*

A la hora de estructurar metodológicamente las predicciones, en particular con respecto al espacio y distribución geográfica de las ocurrencias delictivas, un concepto utilizado en la literatura es el de *hotspots* – literalmente traducido como "sitios calientes". Éstos corresponden a áreas, de tamaño geográfico diverso, que poseen una alta probabilidad de ocurrencia del crimen [18].

En la Figura 2.3 puede apreciarse el mapa de una ciudad dividido en grillas cuadradas, de las cuales algunas han sido clasificadas como *hotspots* por un modelo de predicción del crimen.

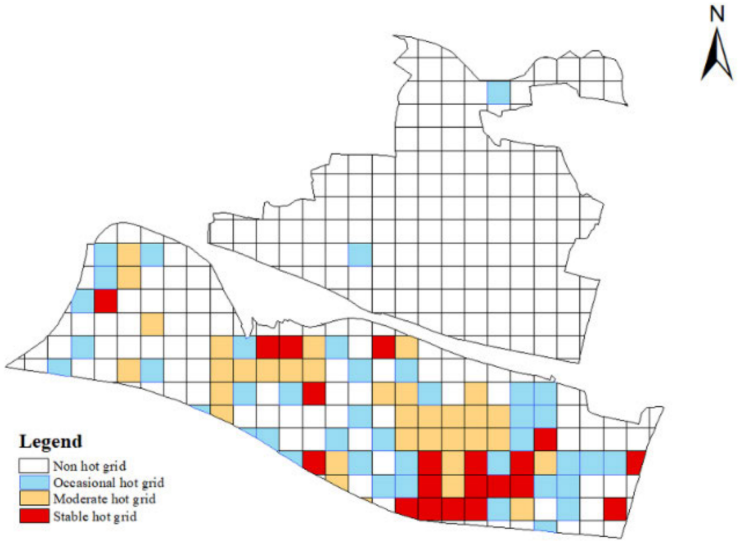

Figura 2.3: Ejemplo de visualización en un mapa grillado [19]

Se pueden observar en la literatura de predicción del crimen tres principales maneras de modelar espacialmente los conjuntos de datos – de tipo espaciotemporal – para el desarrollo de modelos de predicción del crimen.

2.2.1.1. Grillas

El modelamiento según grillas es una técnica que consiste en dividir en subsecciones geométricas los datos espaciales disponibles. Usualmente los datos tienen que presentar latitud y longitud para poder clasificarlos en cada una de las grillas modeladas.

Las grillas suelen darse de manera arbitraria según las necesidades de los investigadores. Estas necesidades pueden incluir asignación de recursos policiales, asignación de recursos pú-

blicos, o el aumento de la eficacia de los modelos de predicción, entre otras razones. Puede proponerse una única configuración de grillas, o puede proponerse un conjunto de configuraciones, para luego después probar cuál es la configuración más exitosa según la pregunta de investigación a resolver.

Las grillas pueden ser cuadradas, circulares o, según tendencias recientes, hexagonales [20]. Sus dimensiones suelen expresarse en números enteros, y en metros, pies, o en unidades de latitud y longitud. Ésta última manera de definir las grillas suele ser más sencilla en términos de programación, dependiendo de la naturaleza de los datos disponibles.

2.2.1.2. Análisis de densidad

Existen estudios en los cuales el foco no está mayormente ubicado en el modelamiento espacio−temporal del conjunto de datos, sino que en el desarrollo y perfeccionamiento de modelos de aprendizaje de máquinas, y de sus resultados predictivos. Por tanto, en vez de utilizar recursos para probar distintas configuraciones de grillas, o simplemente para no enfocarse en definir manualmente un tamaño adecuado, se realizan análisis de densidad.

La técnica más frecuentada por los investigadores, cuando se da este caso, es Kernel Density Estimator [21]. A grandes rasgos, esta técnica agrupa todos los puntos del conjunto de datos en un mapa y define zonas en las cuales las densidades son mayores o menores. Esta manera de modelar conjuntos de datos permite agrupar observaciones sin una forma geométrica definida.

2.2.1.3. Agrupación distrital

En numerosas ocasiones los datos disponibles para los modelos de predicción del crimen incluyen como una de sus variables la pertenencia a alguna zona dentro de la ciudad, tales como distritos, comunas, o regiones. Se observa también en la literatura la agrupación de los conjuntos de datos en estas zonas predefinidas para diversos fines predictivos.

2.2.2. Tipos de modelos de predicción del crimen

Desde el enfoque del desarrollo de modelos de aprendizaje automático para la predicción del crimen, hay dos enfoques mayormente utilizados en la literatura, los modelos basados en regresiones y aquellos basados en clasificación, tanto binaria como multiclase.

Cuando se abordan problemas de predicción del crimen mediante modelos de regresión, lo que se busca predecir es un número o valor, en la mayoría de los casos continuo, que pudiera estar relacionado al número de crímenes ocurridos en un cierto sector, en un determinado momento. En estos caso, la pregunta de investigación a responder podría verse como "¿Cuántas ocurrencias de actos criminales habrá en un próximo período, en una región determinada?".

Por otra parte, existen también los modelos de predicción del crimen basados en modelos de clasificación binaria o multiclase. En el primero de estos casos, la pregunta de investigación podría verse como "¿Habrá o no un crimen, o un acto delictivo, en un siguiente período determinado, para una región en específico?". Para el caso multicategoría, en cambio, suele

verse en la literatura que se busca predecir en específico el tipo de crimen que ocurrirá, entre otros diversos problemas de multietiquetado.

Según una investigación propia, del total de publicaciones científicas que contienen al menos un modelo de predicción del crimen basado en aprendizaje de máquinas, las cuales fueron contabilizadas en 190 (hasta fines de 2021), un $45,65\,\%$ contiene modelos de regresión, un $33,7\,\%$ contiene modelos de clasificación binaria, un $15,22\,\%$ contiene modelos de clasificación multiclase, y un $5,43\,\%$ contiene modelos que escapan de esta categorización, o, dicho de otro modo, son modelos orientados a resolver un problema o caso de uso sumamente específico para fines de cada investigación. Una visualización de esta distribución puede verse en la Figura 2.4.

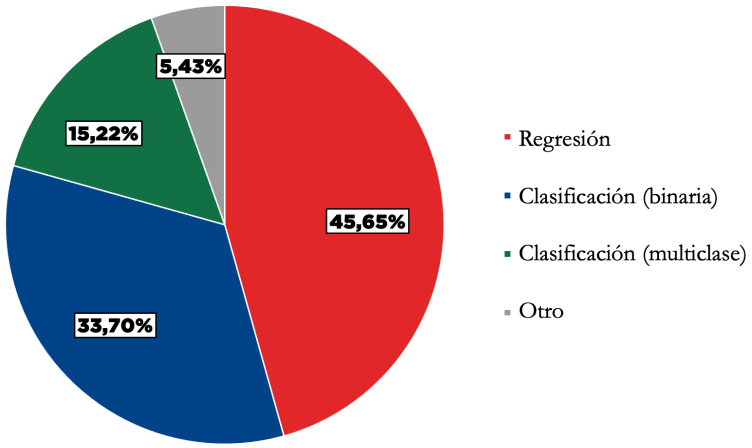

Figura 2.4: Tipo de modelos de predicción del crimen, según el tipo de variable a predecir [elaboración propia]

2.2.3. Tipos de datos utilizados

Los modelos de predicción del crimen son usualmente trabajados por los investigadores en conjunto con las partes interesadas en la prevención del delito, tales como son los departamentos de policía, las municipalidades, gobiernos locales, fiscalías, entre muchos otros actores involucrados. Por tanto, es común que los problemas de predicción del crimen abordados en la literatura cuenten con datos de criminalidad obtenidos de primera fuente sobre crímenes ocurridos en un cierto sector para un período determinado de tiempo. Estos datos de crímenes suelen ser datos atómicos de observaciones de delitos, acompañados de una fecha, hora y lugar de su ocurrencia, entre otras variables.

Según una recopilación propia, el $90\,\%$ de las publicaciones científicas que presentan un modelo de predicción del crimen basan su análisis en datos de tipo espaciotemporales. Se cuenta con una ubicación espacial de un crimen cometido, ya sea latitud y longitud, o una descripción del lugar; el nombre de la comuna, provincia, región, o barrio, entre otros. Por el lado del modelamiento temporal, se suele contar con una ventana horaria de ocurrencia del crimen.

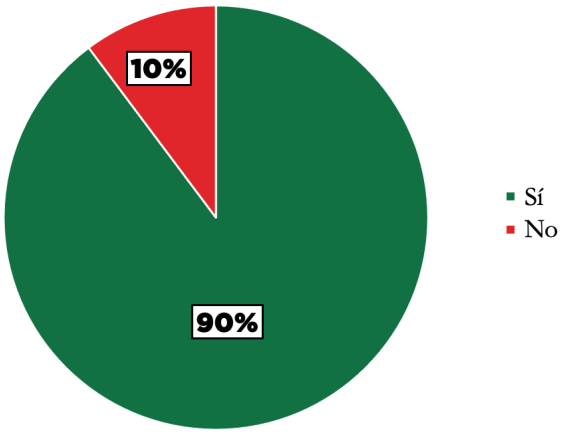

Figura 2.5: Publicaciones que utilizan, como mínimo, variables de tipo espaciotemporal para la generación de modelos predictivos [elaboración propia]

Por otra parte, los estudios de predicción del crimen pueden complementar también el conjunto de datos original provisto por una de las partes interesadas ya mencionada con conjuntos de datos complementario, para, entre otras razones, afinar la precisión de las predicciones, y abordar más factores que pudieran resultar como predictores del delito. Sin embargo, esto puede ser un aporte exclusivamente dependiendo del caso de estudio específico a abordar en cada caso.

2.2.4. Estado del arte y contribución

La predicción del crimen no es una necesidad reciente, sin embargo, es un campo reciente de estudio, en particular en el campo de la ciencia de datos y el aprendizaje de máquinas. Analíticas prescriptivas de este tipo han ganado tracción durante los últimos años tanto en investigación como en la práctica [22].

Según una investigación propia, inédita hasta la fecha, y llevada a cabo en paralelo al presente trabajo investigativo, el 87 % de todas las publicaciones científicas que contienen al menos un modelo de aprendizaje automático de predicción del crimen han sido publicadas durante los últimos cinco años. En esa misma línea, estudios recientes han probado que este tipo de investigaciones pueden llegar a alcanzar un gran desempeño, tanto en capacidad predictiva, como en términos de la implementación de los modelos [23].

Diversos métodos de aprendizaje de máquinas se han utilizado para el desarrollo de modelos predictivos del crimen. Uno de los primeros acercamientos fue el de Olligschlaeger, en 1997 [24], quien desarrolló la primera red neuronal para la predicción del crimen, consistente en nueve neuronas, y sólo una capa escondida, debido a las limitaciones computacionales de la época. Su modelo consistía en un predictor multiclase para una grilla definida de 445 celdas.

Según una investigación propia, de la totalidad de publicaciones científicas de predicción del crimen que contienen al menos un modelo de aprendizaje de máquinas, aproximadamente un 43,3 % contiene un modelo basado en redes neuronales y aprendizaje profundo.

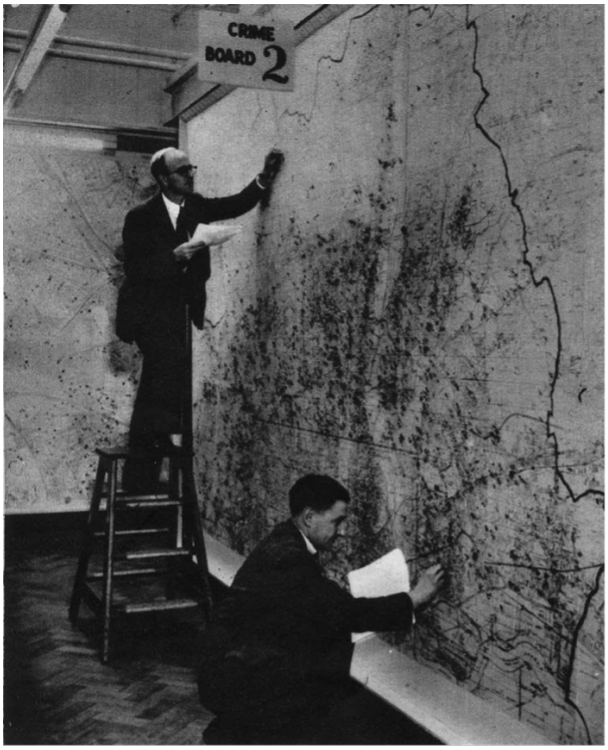

Figura 2.6: Dos policías registrando crímenes en un mapa, 1947 [25]

Modelos más complejos requieren herramientas más sofisticadas, por lo que es común que modelos de mayor complejidad utilicen técnicas tales como $LSTM$ y redes convolucionales para el estudio de imágenes satelitales y generación de mapas grillados.

A pesar de esto, y según una revisión propia, más de la mitad de los modelos de predicción del crimen publicados en la literatura no utilizan redes neuronales artificiales. En este campo de estudio, se trata de un *tradeoff* de importancia, ya que a mayor desempeño de un modelo, menor comprensibilidad poseen [26]. A pesar de que se han desarrollado contraejemplos interesantes en investigaciones recientes [27], es un factor importante a considerar en los modelos de ciencia de datos, en particular, en aquellos de datos tan sensibles para la seguridad pública y ciudadana como los que en este estudio se presentan.

Existen estudios previos, para el Gran Santiago, que han trabajado el patrullaje preventivo mediante herramientas de aprendizaje de máquinas, utilizando datos de Carabineros de Chile, con un foco en la información de los turnos del personal, entre otras variables [28]. El presente estudio, más que enfocarse en el diseño de arquitecturas computacionales a la vanguardia para la predicción y clasificación del delito, y en un trabajo colaborativo con las fuerzas policiales, busca posicionarse como una implementación novedosa de herramientas clásicas de clasificación binaria para el conjunto de datos disponible, enfocándose en el modelamiento metodológico y escalable de las observaciones, con el fin último de la predicción del crimen, según los objetivos y alcances que presenta el acercamiento al problema. Un análisis exhaustivo de la literatura pertinente no permitió dar con ningún estudio que utilizara reportes ciudadanos directos e inmediatos, mediante una plataforma móvil de las características de SOSAFE, para la prevención y predicción del delito.

2.3. Aspectos éticos

Durante los últimos años, los modelos de predicción del crimen han sido motivo de atención en la literatura, la academia y la opinión pública por, entre otras cosas, despertar dudas sobre la legitimidad de sus predicciones. Esta es una preocupación que debiese estar siempre presente en la gran mayoría de los modelos de ciencia de datos, en mayor o menor grado, dependiendo de la naturaleza del problema en cuestión.

Al ser los modelos de predicción del crimen modelos de ciencia de datos que se enmarcan en las políticas públicas, y dado que su implementación afecta directamente a los ciudadanos (con particular correlación con la distribución socioeconómica en el territorio), éstos son particularmente sensibles a este análisis ético, ya que podrían llegar a perpetuar sesgos ya existentes dentro de la sociedad [29].

2.3.1. Sesgos presentes en el conjunto de datos

El sesgo, en particular el sesgo estadístico, se define como un error sistemático el cual es producido, en la mayoría de los casos, por un muestreo desbalanceado de la población que pretende representar, y que entrega resultados que podrían diferir sustantivamente de la realidad que pretende modelar [30].

Los conjuntos de datos presentan, a nivel general y en mayor o menor magnitud, sesgos propios, ocasionados tanto por la recolección de los datos como por el desarrollo de los modelos, entre otras razones. Este es el caso en particular del conjunto de datos provisto por SOSAFE, el cual se detallará en la siguiente subsección, y en el capítulo que le sigue.

Los sesgos para un problema de ciencia de datos pueden abordarse de diversas maneras, según sean los diversos factores involucrados. Algunos de los sesgos para este caso de estudio en particular no pueden abordarse de manera sencilla, por lo que sólo a continuación se plantean y, en general, se suponen, sin probarse empíricamente necesariamente, pero acompañados de una argumentación pertinente. La discusión y análisis de los sesgos es crucial antes de trabajar cualquier conjunto de datos, dado que afecta el desarrollo e implementación de los modelos desde su concepción.

Para el caso de SOSAFE, es trivial establecer que los usuarios de la aplicación no se encuentran homogéneamente distribuidos a lo largo y ancho de la Región Metropolitana (la región de estudio en cuestión) ni mucho menos a lo largo y ancho de Chile. Quienes descargan la aplicación lo hacen porque, en primer lugar, poseen las herramientas para descargarla, tales como, por ejemplo, un teléfono móvil. Probablemente poseen la capacidad de leer y escribir, además de una conexión a internet lo suficientemente estable para permitirles hacer uso de la aplicación. Además, tienen el interés y deseo de reportar actos de vandalismo a nivel vecinal, o sencillamente de realizar reportes de vida comunal en general, lo cual no es necesariamente transversal a todos los ciudadanos. Estas características, particularmente las primeras, implican una división socioeconómica heterogénea inherente, siendo una de las principales variables a considerar la distribución territorial de los usuarios (distribución heterogénea que, de hecho, es posible corroborar al observar los datos disponibles).

Además de las características propias de los usuarios, los reportes en sí mismos pueden presentar sesgos propios de su recolección. Éstos se ven enmarcados en una situación a nivel país que se escapa de lo ordinario. Los reportes fueron recolectados entre abril de 2019 y abril de 2021, por lo que hay dos momentos importantes en la historia reciente del país en los cuales se ven involucrados; la pandemia sanitaria del Covid-19, y la revuelta social de octubre del 2019. No necesariamente un modelo de ciencia de datos construido según la información que aporta un período excepcional de tiempo será escalable y proyectable a más épocas.

Más allá del sesgo histórico, los reportes funcionan para efectos de este proyecto como una estimación de la criminalidad real, la cual es sumamente difícil o imposible de medir con precisión óptima. Por tanto, los resultados que entreguen los modelos sólo serán un acercamiento aproximado a la medición de la criminalidad real, de acuerdo al detalle según ésta se quisiera medir. Se puede hablar de un estimador sesgado del delito.

Uno de los aspectos en los cuales los reportes del crimen se diferencian con respecto a la ocurrencia del acto criminal en sí, es la diferencia horaria en la cual ambos ocurren. Un reporte es, en la gran mayoría de los casos, un informe posterior a la realización del crimen en sí, y la medición de la diferencia horaria entre el acto y su respectivo reporte, según la categoría a la que el reporte pertenece en sí, escapa de los alcances del presente estudio.

A nivel espacial, por su parte, se debe trabajar bajo el supuesto de que la diferencia de ubicación de un reporte con el hecho del acto criminal en sí no es muy grande. En otras palabras, se trabajará bajo el supuesto de que la latitud y longitud registradas en cada reporte están fuertemente relacionadas a la ocurrencia del acto criminal en sí. Este es un supuesto importante, y a pesar de que no se cumplirá en la totalidad de los reportes disponibles, se procederá con el caso de investigación bajo la premisa de que, incluso en el caso de mayor diferencia, ésta no escapará en mayor grado de, al menos, los límites comunales.

Finalmente, un último sesgo inherente al conjunto de datos digno de tener en consideración es la categorización manual a la cual son sometidos los reportes, luego de su escritura, por parte de los usuarios. Los usuarios contaban con total libertad por parte de la aplicación para etiquetar sus reportes en cualquier categoría que les pareciera pertinente. Es más, durante la ventana temporal en la cual se recopilaron los reportes disponibles, la aplicación contaba con una categoría "por defecto", la cual es 'Actividad sospechosa'. Por tanto, es posible afirmar, con un alto grado de certeza, que existen sesgos en el conjunto de datos a nivel de categorías, y que este sesgo es mayor o menor dependiendo de la categoría.

2.4. Caracterización de la empresa

SOSAFE es una empresa chilena fundada en 2014. A la fecha, declaran tener más de 1,5 millones de usuarios en Chile y el mundo.

En su página web [31] declaran ser una aplicación de carácter vecinal. Indican: "Aplicación vecinal: la nueva comunicación entre vecinos y municipios. Nuestra galardonada aplicación móvil ayuda a la gente a conectar con sus vecinos, como también notificar a las autoridades sobre temas en sus barrios y emergencias de seguridad". La empresa manifiesta una posición estratégica – es decir, a largo plazo, que la caracteriza frente a sus usuarios y que la guía en la toma de decisiones de negocio – relativa a la seguridad y a la cercanía vecinal.

Un modelo de negocios describe las bases sobre las que una empresa crea, proporciona y capta valor [32]. El modelo de negocios que la compañía presenta podría catalogarse como, en primer lugar, *B2B* – *business-to-business*; la organización ofrece sus servicios a otras empresas. En el caso de SOSAFE, la empresa tiene planes y servicios para municipalidades y empresas. La empresa declara tener experiencia con clientes como Walmart, Pedidos Ya y la Municipalidad de Lo Barnechea, por mencionar algunos [31]. Sin embargo, por otra parte también podría catalogarse como una compañía *B2C*, es decir, que ofrece sus servicios a personas, bajo una modalidad y foco diferentes. El producto dirigido a usuarios es su aplicación móvil.

Los datos disponibles son reportes emitidos por los usuarios de SOSAFE, a lo largo y ancho de todo el país, entre los años 2019 y 2021. Estos reportes son escritos de manera manual por los usuarios, a los cuales posteriormente les asignan una categoría a la que pertenecen.

Figura 2.7: *Landing page* de SOSAFE [31]

2.5. Contexto geográfico y división político-administrativa

Chile es una nación sudamericana, que se divide en dieciséis regiones administrativas. Limita territorialmente con las naciones vecinas de Argentina, Perú y Bolivia.

Su territorio es más largo que ancho, extendiéndose por 4.270 $[km]$ desde el norte hasta el sur. El ancho máximo del país son 445 $[km]$, en los 52°21' S (Región de Antofagasta), y el mínimo son 90 $[km]$, en los 31°37' S (Región de Aysén) [33].

La Región Metropolitana de Santiago es la región más habitada de Chile, con una población de 7.112.808 personas. Posee 6 provincias y 52 comunas, de las cuales 18 son rurales. Tiene una superficie total de $15.403,2$ $[km^2]$ [34]. Se ubica entre los 32º 55' y 34º 19' de latitud sur, y entre los 69º 47' y 71º 43' longitud oeste [35].

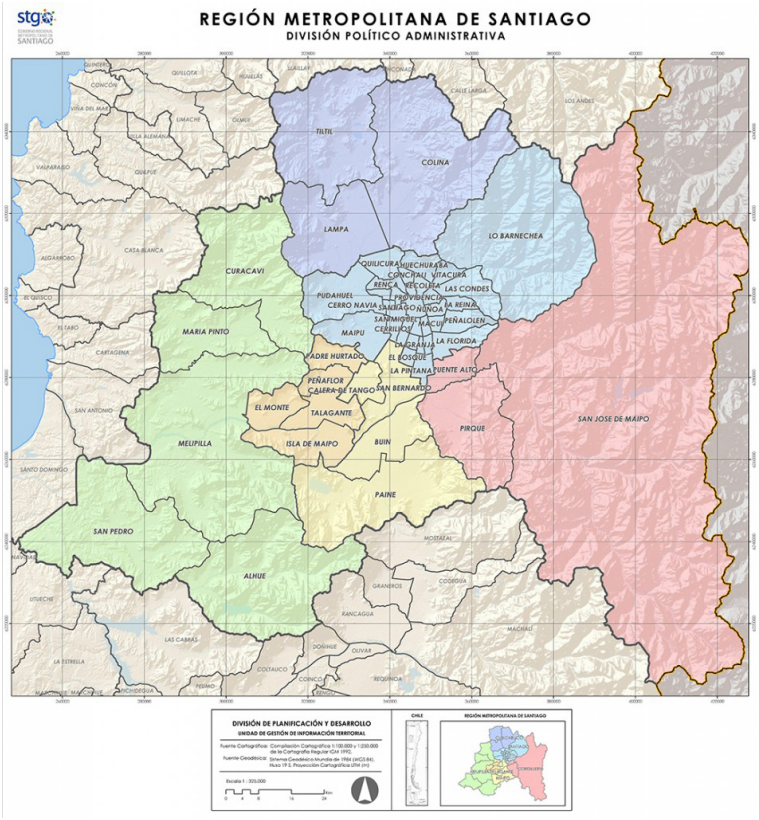

Figura 2.8: División político–administrativa de la Región Metropolitana [34]

En el centro se ubica la provincia de Santiago, compuesta por 32 comunas. Cuenta con el 76 % de la población regional, y es la provincia con la mayor densidad poblacional del país. Por su parte, el área metropolitana de Santiago se encuentra conformada por el Gran Santiago, que corresponde a una conurbación de cuarenta comunas, más seis zonas periurbanas, entre las cuales se cuentan Colina, Lampa, Buin, Peñaflor, Talagante y Paine [36].

El nombre de todas las comunas de la Región Metropolitana, junto con la provincia a la que pertenecen, se encuentran listados en la Tabla A.1 de Anexos.

Capítulo 3

Metodología

3.1. Análisis exploratorio de datos

Los datos disponibles son reportes emitidos por los usuarios de SOSAFE, a lo largo y ancho de todo el país, entre los años 2019 y 2021.

Los reportes son escritos por los usuarios, quienes les asignan una categoría. El conjunto de datos provee, además del *input* manual de cada usuario, la fecha y hora exacta en la cual fue publicado cada reporte, además de la ubicación geográfica exacta del lugar de su emisión.

Son 1.789.486 reportes totales. Los atributos exactos del conjunto de datos, junto con una descripción y caracterización de cada uno, se listan a continuación:

- **ID**: Identificador único, en formato de texto – *string*.
  Ejemplo "73b3f865-238f-58f2-ac1e-24b8b2541990"
- **Descripción**: Reporte de los hechos. Formato texto – *string*.
  Ejemplo: "Grupo de 3 a 5 personas en estado de ebriedad"
- **Categoría**: Registro de la categoría a la que se asocia cada reporte. En formato de texto – *string*.
  Ejemplo: "Mascota Perdida"
- **Fecha y hora**: Fecha y hora exacta del momento de publicación del reporte. En formato *datetime*.
  Ejemplo: "2019-04-26 21:58:42"
- **Latitud**: Latitud exacta de la publicación del reporte en la plataforma. En formato numérico – *float*.
  Ejemplo: "−33.42681"
- **Longitud**: Longitud exacta de la ubicación de la publicación del reporte. En formato numérico – *float*.
  Ejemplo: "−70.16854"
- **Ubicación**: Descripción manual e individual del lugar de ocurrencia de cada reporte enviado. En formato de texto – *string*.
  Ejemplo: "Irarrázaval #2335, Ñuñoa"

3.1.1. Categorías presentes en el conjunto de datos

Los reportes cuentan con amplia diversidad en su contenido. Las categorías presentes en los datos no son sólo asociadas a criminalidad, sino que también a vida vecinal y barrial de todo tipo. Asimismo, no todos los reportes pertenecen conceptualmente a la categoría bajo la cual son registrados, lo cual supone una dificultad para el manejo de los datos.

Algunos ejemplos de reportes, junto con la categoría a la que pertenecen, pueden observarse en la Tabla 3.1.

Tabla 3.1: Ejemplo de reportes

Contenido del reporte	Categoría
"Acaban de robar un vehículo Nissan NAVARA color Rojo ppu: FK-L*-**"	Robo auto
"Nuestra Pimienta sigue extraviada, si la ves por favor llamar al +5697603****"	Mascota perdida
"Corte de luz (Centenario)"	Aviso comunitario
"Ruidos molestos pasaje adriana cousiño 3** ventanas abiertas no dejan dormir"	Vandalismo
"Música y voces muy fuertes en Antillanca 9*** 2do piso. Al parecer depto 22"	Robo casa

Hay 37 categorías distintas presentes en los reportes emitidos en la Región Metropolitana. Hubo 35 categorías distintas presentes en el período 2019 – 2020, y 37 distintas en el período 2020 – 2021. Las categorías presentes en la Región Metropolitana, junto con la cantidad de reportes para cada una, pueden observarse en la Tabla 3.2.

Hay 55 categorías distintas presentes a lo largo de todo el país. De las 18 categorías a nivel nacional que no están presentes en la Región Metropolitana, gran parte son categorías relacionadas a minería. Un análisis de los datos permite concluir que la mayoría de estos reportes fueron emitidos en la Región de Antofagasta.

Como puede observarse en la Tabla 3.1, al asignar los usuarios manualmente la categoría a la que pertenece cada reporte descrito, esto puede derivar en que los reportes no queden todos adecuadamente categorizados, lo que implica un sesgo inherente al conjunto de datos (ver sección 2.3.1).

Las categorías presentes a nivel nacional, junto a la cantidad de reportes para cada una de ellas, se encuentra en la Tabla A.1 de Anexos.

Tabla 3.2: Cantidad total de reportes disponibles, según categoría, en la RM

Categoría	Cantidad de reportes
Seguridad	271.498
Vandalismo	180.626
Aviso comunitario	173.860
Mascota perdida	150.091
Ruido molesto	138.297
Actividad sospechosa	120.858
Prueba	99.402
Vender	60.830
Otros públicos	46.504
Comercio habilitado	38.245
Alumbrado público	33.343
Buena acción	27.751
Bomberos	24.964
Accidente	17.602
Vehículo abandonado	16.844
Ambulancia	16.249
Comercio ambulante	15.158
Basura	15.046
Semáforo defectuoso	13.267
Robo auto	13.139
Robo a persona	11.709
Comprar	6.871
Alcantarilla sin tapa	6.341
Robo casa	5.604
Poda de Árboles	5.372
Fuga de agua	4.689
Señalética	2.952
Incumplimiento del toque de queda	2.459
Grafitis	1.228
Situación de calle	743
Pavimento Dañado	693
Paradero en mal estado	532
Sistema eléctrico en mal estado	254
Vereda en mal estado	72
Quema ilegal	68
Animal en vía	58
Robo de instalaciones	57

3.1.2. Distribución espacial de los reportes

De los 1.789.486 reportes totales, el $90,3\,\%$ pertenecen a la Región Metropolitana.

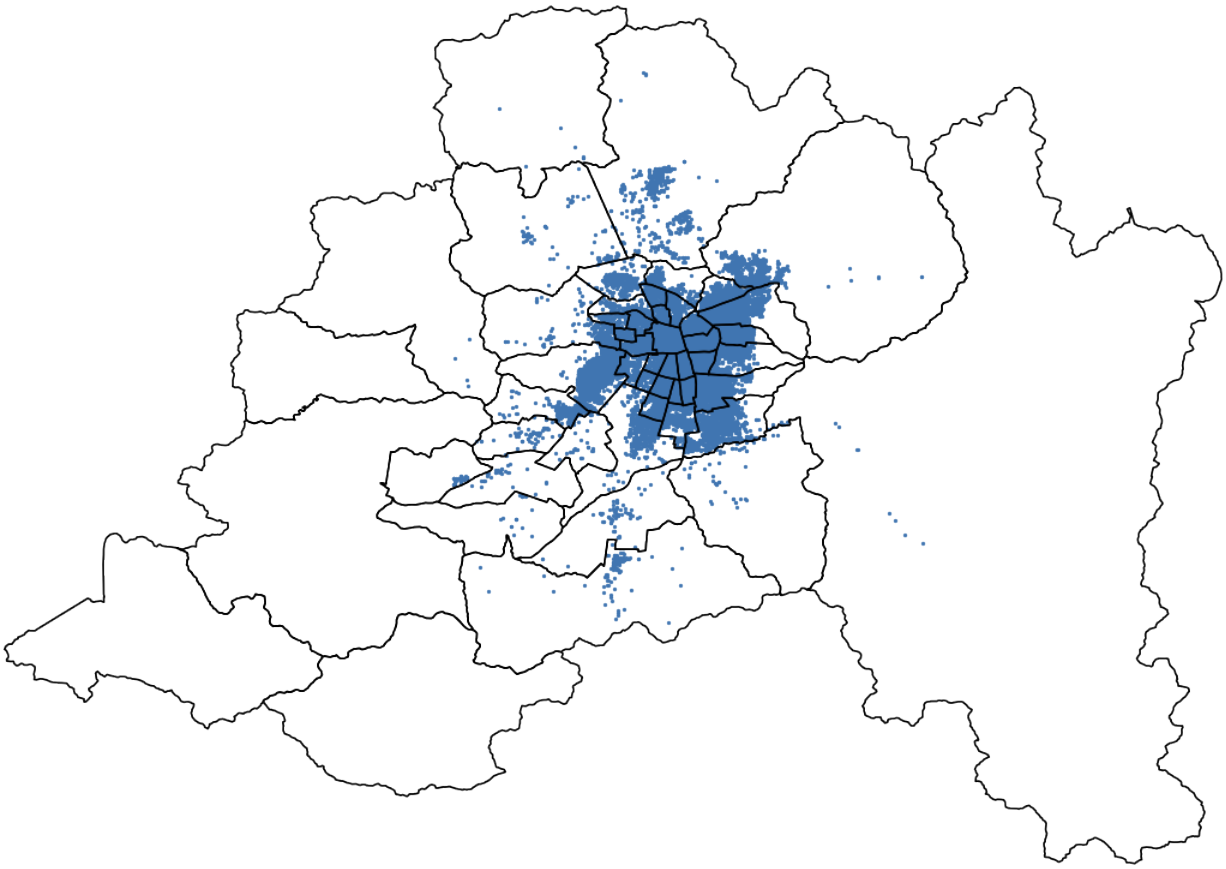

Figura 3.1: Reportes emitidos en la Región Metropolitana durante el mes de septiembre de 2019 [37]

Existen otras regiones del país que también concentran gran cantidad de observaciones. Estas, en su gran mayoría, reflejan la presencia de otras ciudades importantes, que, sin embargo, no alcanzan una gran relevancia en el conjunto de datos en comparación a la Región Metropolitana. Por alcances del problema de investigación, estos otros puntos, de importante densidad, no serán estudiados en mayor detalle.

Una aproximación a la distribución de los datos disponibles a nivel nacional, para un mes en particular, puede observarse en la Figura B.1 de Anexos.

3.1.3. Distribución temporal de los reportes

El primer reporte disponible corresponde al 26 de abril de 2019. El último reporte disponible corresponde al 26 de abril de 2021. Son dos años exactos de observaciones disponibles.

En la Figura 3.2 puede observarse la distribución temporal de la cantidad de reportes, a nivel nacional. Dos principales conclusiones pueden extraerse.

1. La primera es que hubo un aumento en la tendencia del uso de la aplicación a lo largo del tiempo, existiendo una menor cantidad de reportes diarios, en promedio, para los primeros meses de datos disponibles, con respecto a los últimos.

2. La segunda principal conclusión que puede observarse en el histograma es el radical aumento en el número de reportes para la tercera semana de octubre de 2019. Este evento coincide con el comienzo de las masivas manifestaciones en Chile observadas durante aquel mes.

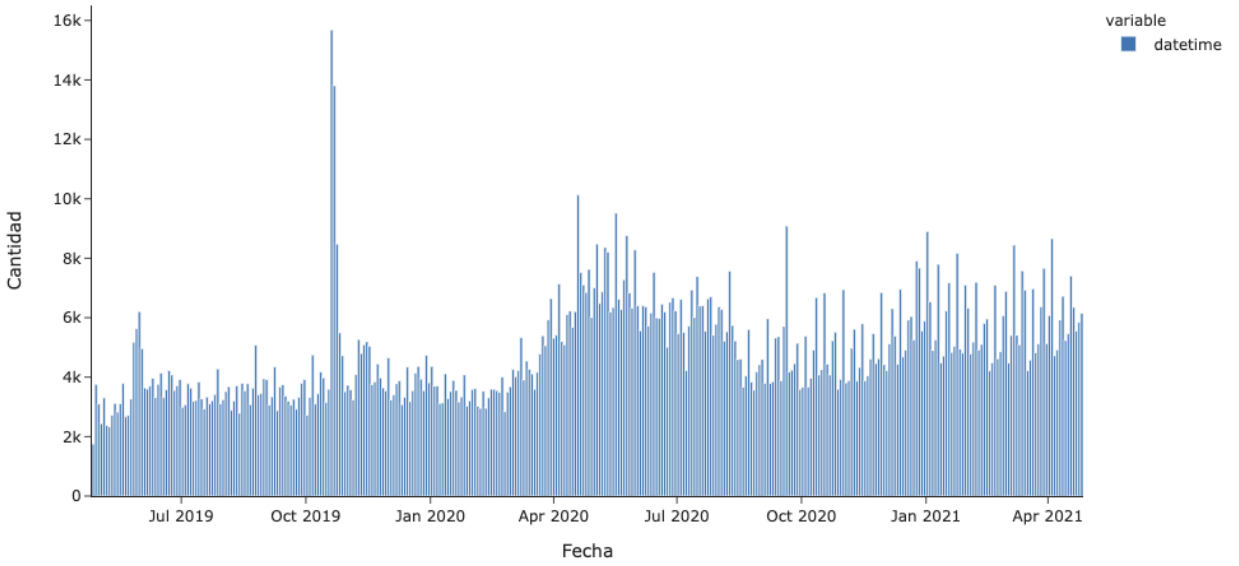

Figura 3.2: Cantidad de reportes según fecha

3.1.4. Calidad de los datos

Los atributos 'ID', 'Fecha y Hora', 'Latitud', 'Longitud' y 'Categoría' solo cuentan con una observación nula cada una, que corresponde a la misma observación, que tiene todos sus atributos nulos. Por su parte, 374 observaciones, incluyendo aquella mencionada, poseen el atributo de 'Ubicación' nulo. Por último, la descripción es el atributo que más veces aparece como nulo en los datos originales, presentando esta condición en 102.216 observaciones, un $5,72\,\%$ del conjunto de datos disponible.

3.2. Modelamiento del conjunto de datos

Para responder la pregunta de investigación, y tener la capacidad de determinar cuándo y en qué lugar ocurrirán actos criminales, de las características que se han descrito, es necesario plantear agrupaciones de los datos tanto temporales como espaciales. Esto va en la misma línea de lo que se ha trabajado en literatura que aborda problemas similares, y permite una mayor abordabilidad del problema con respecto a si se trabajaran las observaciones a nivel atómico de los datos.

3.2.1. Modelamiento espacial del conjunto de datos

Los datos se agrupan espacialmente en grillas cuadradas definidas arbitrariamente. El tamaño de las grillas responde a la necesidad de asignar recursos de manera táctica y/o estratégica. Por tanto, se descarta la posibilidad de agrupar los datos en grillas de, por ejemplo, 3 $[km]$ x 3 $[km]$, dada la dimensionalidad de la ciudad, los barrios y cuadras, la cual requiere granularidad más finas. Las dimensiones de cada una de las configuraciones espaciales son, aproximadamente:

1. La **primera configuración espacial** determina grillas de $1,11$ $[km]$ x $0,93$ $[km]$, con un área de $1,03$ $[km^2]$.
2. La **segunda configuración** propuesta define grillas cuadradas de $0,78$ $[km]$ x $0,65$ $[km]$, con un área de $0,505$ $[km^2]$.
3. La **tercera configuración espacial** propuesta determina grillas de $0,56$ $[km]$ x $0,46$ $[km]$, con un área de $0,258$ $[km^2]$.

Se proponen diferentes configuraciones de grillas para poder implementar los modelos en cada una de ellas, con el fin de que los resultados obtenidos para cada una entreguen conclusiones sobre las ventajas y desventajas de cada modelamiento. En comparación, la segunda configuración de grillas posee la mitad del área de la primera configuración, y la tercera, la mitad del largo y ancho. Una discusión más profunda sobre el origen de la elección de estas configuraciones se propone en la subsección 4.3.2.2.

3.2.2. Modelamiento temporal del conjunto de datos

3.2.2.1. Diseño del nuevo conjunto de datos

Para complementar la agregación espacial de los datos, en vista de la pregunta de investigación, se plantean también agrupaciones a nivel temporal.

Se plantean períodos que pudieran resultar de interés, en vista de los resultados a obtener y sus posibles implementaciones. En este sentido, se proponen cuatro granularidades temporales distintas; **dos–semanas**, **semanal**, **tres–días** y **diaria**. Para cada granularidad espacial y temporal se suma la cantidad de reportes que hubo para cada categoría. La intención es utilizar esta cantidad como variable clave para la predicción.

En general, la intuición detrás del modelamiento es descubrir cuántos días hacia atrás bastan para predecir la variable objetivo asociada a la delincuencia, la cual será definida en breve. Por ejemplo, para la configuración de **ventanas temporales semanales**, se suman todos los reportes, para cada categoría, durante la semana previa a la ocurrencia de la variable objetivo a predecir. Esta semana es catalogada como '$i-7$', para la semana i de ocurrencia de la actividad criminal. Se incluyen, en el mismo ejemplo de la configuración semanal, el número de reportes para cada categoría entre la semana previa y la anterior $(i-14)$; los reportes de la semana ocurrida tres semanas atrás $(i-21)$, cuatro semanas atrás $(i-28)$ y un año atrás $(i-365)$. Una caracterización de cómo funciona este modelamiento, para los primeros cuatro períodos propuestos, puede verse en la Figura 3.3.

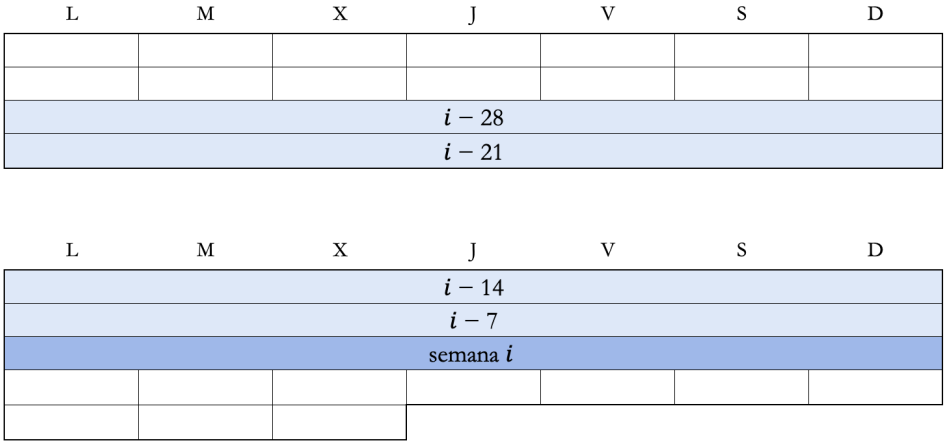

Figura 3.3: Configuración semanal del conjunto de datos

De manera complementaria a la configuración anterior, se propone también una configuración de ventanas temporales de **dos semanas**, con sumas de reportes para cada categoría de catorce días previos $(i-14)$, para el período entre dos y cuatro semanas atrás de reportes $(i-28)$, seis semanas $(i-42)$, ocho semanas $(i-56)$, y un año atrás $(i-365)$. Este es la segunda de las cuatro configuraciones temporales propuestas.

Se proponen dos períodos más, de mayor fineza en la granularidad. El primero de estos es la **granularidad temporal de tres–días**, la cual captura la información tres días antes del período a predecir $(i-3)$, para el período entre tres y seis días antes $(i-6)$, entre seis y nueve $(i-9)$, nueve y doce $(i-12)$ y entre doce y quince días antes $(i-15)$. A esta configuración se agregan de manera complementaria variables para capturar estacionalidades para una $(i-7)$, dos $(i-14)$ y cuatro semanas previas $(i-28)$.

Finalmente, se propone una cuarta granularidad temporal, esta vez, de ventanas temporales de un día. La **granularidad diaria** incorpora predictores para todos los días de la semana previa a la predicción esperada $(i-1,\ i-2,\ i-3,\ i-4,\ i-5$ e $i-6)$, además de las ocurrencias de una $(i-7)$, dos $(i-14)$ y cuatro semanas previas $(i-28)$, además de un año exacto hacia atrás $(i-365)$.

3.2.2.2. Construcción del nuevo conjunto de datos

En el modelamiento y construcción de los nuevos conjuntos de datos utilizados para la predicción, las observaciones del último año disponible de datos ($2020-21$) toman una mayor preponderancia con respecto a sus símiles del primer año ($2019-20$).

Esto ocurre por dos principales razones. La primera es que hay más reportes disponibles para el segundo año con respecto al primero (ver Figura 3.2). La segunda es que, dado el diseño del conjunto de datos propuesto, en particular para su componente temporal, sólo las observaciones pertenecientes al último año de datos tienen la capacidad de recurrir a información de reportes que ocurrieron un año atrás, a diferencia de los reportes del año 2019, por ejemplo, que no tienen datos para 365 días atrás, y en algunos casos borde, ni siquiera para algunos días o semanas previas.

Por las razones planteadas, la construcción del conjunto de datos a utilizar ocurre en dos partes. Primero, se separan los datos en dos mitades. La primera mitad de los datos comprende desde el 29 de abril del 2019 hasta el 26 de abril de 2020, capturando exactamente un año calendario de observaciones (el 29 de abril de 2019 fue un lunes y el 26 de abril de 2020 fue un domingo). El segundo conjunto de datos comprende observaciones desde el 27 de abril de 2020 hasta el 25 de abril de 2021, abarcando así también un año calendario exacto de observaciones, Este acercamiento a la construcción del conjunto de datos facilita el modelamiento posterior, así como el planteamiento conceptual del problema.

Posterior a lo planteado, el foco para la predicción es el segundo conjunto de datos construido. El primero le sirve como *input*, y de ahí se sacan las filas de datos necesarias para construir los predictores vinculados a períodos pasados de tiempo.

Dependiendo de la granularidad espacial y temporal de cada uno de los casos planteados a estudiar, las grillas espaciales definidas se repiten numerosas veces a lo largo del nuevo conjunto de datos, según la granularidad temporal definida. Por ejemplo, para el caso semanal, cada una de las grillas definidas en la Región Metropolitana se repite al menos cincuenta y dos veces. Dependiendo de la fineza de la granularidad en cada caso, esto puede generar que grillas en las que nunca pasa nada, a nivel de reportes en SOSAFE, se vean sobrerrepresentadas en el conjunto de datos una numerosa cantidad de veces. Este profundo desbalance de los datos se aborda posteriormente (ver sección 3.2.6.4).

En la Tabla B.1 de Anexos puede encontrarse una caracterización de cada una de las granularidades propuestas, diferenciando por modelamiento espacial y temporal, e indicando el balanceo de la clase positiva de los datos, es decir, aquellas observaciones que presentan actividades criminales que posteriormente serán implemento vital para el aprendizaje de los modelos.

3.2.3. Categorías

La agregación temporal para la generación de los conjuntos de datos, la cual es el principal *input* para los modelos de aprendizaje de máquinas, implica repetir, al menos cinco veces, todas las categorías, como nuevas columnas del conjunto de datos formado. Con el fin de reducir el número de categorías, se proponen once nuevas macrocategorías, que agrupan temáticamente a las categorías originales.

Para construir las nuevas macrocategorías, se estudió la semejanza que tenía el contenido de los reportes entre sí. Consistió en un extenso trabajo iterativo, con distintas agrupaciones propuestas, cuya propuesta definitiva se presenta en la Tabla 3.3.

Para ejemplificar la agrupación temática, puede verse en la Tabla A.3 de Anexos que la macrocategoría Minería corresponde a la agrupación de ocho de las categorías originalmente provistas, entre las cuales se cuentan 'Derrame', 'Evento climático' y 'Herramienta en mal estado', entre otras. Pese a que por sí sola la categoría (por ejemplo) 'Herramienta en mal estado' pudiera no necesariamente estar relacionada a Minería, es posible observar, en el conjunto de datos provisto, que los reportes pertenecientes a esta categoría no son muchos (184, un $0,01\,\%$ del total de reportes disponibles), y están todos concentrados en la Región de Antofagasta, donde existe una importante actividad minera a nivel nacional. Al estudiar el contenido de los reportes, es posible descubrir que están asociados temáticamente a la actividad minera, donde se relacionan de gran manera con las otra siete categorías temáticamente cohesionadas. Así, los reportes agrupados en esta macrocategoría están altamente correlacionados en número de reportes, ubicación geográfica, y contenido temático de los reportes mismos. Este análisis se replica para cada una de las once nuevas categorías temáticas.

Las macrocategorías modeladas se listan a continuación:

1. **Emergencia**
2. **Comercio**
3. **Aviso Comunitario**
4. **Mascota Perdida**
5. **Robo**
6. **Prueba**
7. **Actividad Sospechosa**
8. **Otros Públicos**
9. **Seguridad**
10. **Ruido Molesto**
11. **Minería**[1]

[1] No hay ningún reporte perteneciente a esta categoría en la Región Metropolitana.

Tabla 3.3: Nuevas categorías propuestas en la Región Metropolitana

Categoría	Nueva categoría propuesta
Accidente Ambulancia Bomberos Quema ilegal	**Emergencia**
Vender Comercio habilitado Comercio ambulante Comprar	**Comercio**
Buena acción Aviso comunitario	**Aviso Comunitario**
Mascota perdida Animal en vía	**Mascota Perdida**
Robo auto Robo a persona Robo casa Robo de instalaciones	**Robo**
Prueba	**Prueba**
Situación de calle Vandalismo Actividad sospechosa Vehículo abandonado Grafitis	**Actividad Sospechosa**
Otros públicos Alumbrado público Basura Semáforo defectuoso Alcantarilla sin tapa Pavimento Dañado Paradero en mal estado Sistema eléctrico en mal estado Vereda en mal estado Poda de Arboles Fuga de agua Señalética	**Otros Públicos**
Seguridad	**Seguridad**
Ruido molesto Incumplimiento del toque de queda	**Ruido Molesto**

3.2.4. Determinación de la variable objetivo

El objetivo de los pasos seguidos hasta ahora es, a grandes rasgos, el diseño y la construcción de un conjunto de datos que permita realizar una estimación apropiada, con un nivel de precisión significativo, sobre los actos de delincuencia en la Región Metropolitana, tal como se han caracterizado hasta ahora. Se abordan como actos delictivos todos aquellos que están categorizados en SOSAFE bajo una de las cuatro categorías relativas a robos, las cuales se han posteriormente agrupado bajo una macrocategoría llamada simplemente 'Robo'. Todo acto delictual que no ha sido categorizado como tal ha sido excluido de la variable objetivo a predecir, y el cálculo del sesgo que esto provoca, o la incorporación posterior de reportes a alguna categoría de pertinencia delictual escapa de los alcances del presente trabajo.

Al definir el problema como uno de caracterización binaria, es necesario encontrar una transformación apropiada del conjunto de datos que permita implementar una variable que tome dos valores, uno referente al crimen, y uno que refiera a su ausencia. En problemas de predicción del crimen basados en modelos de clasificación binaria estudiados en la literatura, a este paso se conoce como la determinación del umbral para la caracterización de los *hotspots*. No existe un valor o característica estándar en la literatura para este tipo de casos, y depende siempre de la naturaleza del conjunto de datos y del problema a resolver.

Para el presente estudio, se define una variable y_{ij}, que se caracteriza en la Ecuación 3.1.

$$y_{ij} = \begin{cases} 1 & \text{si hubo al menos un reporte en la categoría Robo para el período } i \text{ y ubicación } j \\ 0 & \text{si no} \end{cases} \tag{3.1}$$

3.2.5. Filtro por región y estadísticas por comuna

3.2.5.1. Filtro espacial de los datos

El objetivo general del problema de investigación es predecir actividades criminales en la Región Metropolitana en base a los reportes proporcionados por SOSAFE. Al contar con datos a nivel nacional, es necesario filtrar los reportes apropiadamente para trabajar y generar modelos predictivos con únicamente datos de la región determinada. Este acercamiento se torna desafiante al no poseer explícitamente la comuna, provincia o región a la cual pertenece cada una de las observaciones. Sólo se cuenta con la latitud, longitud y descripción de la ubicación de cada reporte.

Una primera aproximación al filtrado espacial de los datos desde la cantidad original del conjunto de datos podría ser fiarse explícitamente de la caracterización espacial de la Región Metropolitana proveída por el Gobierno de Chile. "La Región Metropolitana se localiza en la macrozona central del país, aproximadamente entre los 32º y 34º de latitud Sur y los 69º y 71º de longitud Oeste" [38]. Al contar con esta descripción es posible filtrar, iterativamente y con un grado de precisión de algunos decimales, las observaciones presentes en la región según su latitud y longitud. Sin embargo, este acercamiento no deja de ser sólo una aproximación, ya que al presentar el límite inferior de la región una forma cóncava, algunas observaciones pertenecientes a la Región de O'Higgins, la región limítrofe sur, se ven incluidas en el filtro.

Convenientemente, el tercer campo asociado a la ubicación, la descripción manual, está guardada en los datos de manera tal que se le asocia un sector de la ciudad de manera estandarizada, por tanto que por mucho que la descripción del lugar de un reporte es un ingreso manual según cada usuario, indicando lo que le parezca conveniente para asociar un reporte (como la numeración de una casa, calle y/o sector), estas están siempre asociadas a un sector que se repite a lo largo de todas las observaciones, y que suele ser la comuna. Una pequeña manipulación de los datos permite extraer esta última observación estandarizada, que permite justamente rescatar cada una de las cincuenta y dos comunas de la región, y a grandes rasgos todas las comunas del país en la cual se registró un reporte en el período de tiempo estudiado. Sólo hubo seis sectores, en la Región Metropolitana, que fueron registrados como una comuna, sin serlo; Chicureo (Colina), Batuco (Lampa), El Canelo (San José de Maipo), Valle Grande (Lampa), Huertos Familiares (Tiltil) y Linderos (Buin), las cuales se modifican para calzar con su comuna correspondiente.

Además del filtro por región, se eliminan todas las filas que contengan un dato nulo, excepto en la columna de descripción de los hechos, la cual es la más problemática del conjunto de datos (ver sección 3.1.4), pero a su vez posee la menor ponderación en la predicción de todas ellas, según el modelamiento planteado.

Luego de los nuevos filtros aplicados, el nuevo conjunto de datos posee 1.618.379 filas.

3.2.5.2. Estadísticas por comuna

Se incorporan al conjunto de datos modelado estadísticas propias para cada comuna, aprovechando el identificador comunal construido para cada reporte.

Desde la base de datos del Centro de Estudios de Análisis del Delito, perteneciente a la Subsecretaría de Prevención del Delito del Ministerio de Interior y Seguridad Pública del Gobierno de Chile, se incorpora la tasa cada 100.000 habitantes de casos policiales de Delitos de Mayor Connotación Social, y de Violencia Intrafamiliar, al año 2019, para cada comuna de la Región Metropolitana. El número de casos policiales considera las denuncias de delitos que realiza la comunidad en las unidades policiales, más las detenciones que realizan las policias o ante la ocurrencia de delitos flagrantes [39].

Desde el Sistema Nacional de Información Municipal de la Subsecretaría de Desarrollo Regional y Administrativo, perteneciente al mismo ministerio, se incorporan cinco variables; el presupuesto inicial de gastos municipales (código BPIGM), la disponibilidad presupuestaria municipal por habitante (IADM10), el gasto devengado por Servicios de Aseo, Recolección de Basura y Vertederos (IADM92), el gasto devengado por Servicios de Alumbrado Público (IADM93) y el indicador de pobreza CASEN (ISOC001), ajustados al año 2019, en su valor nominal [40]. El sitio incluye cientos de variables más, de las cuales se escogen las mencionadas por su relación con los reportes, las categorías disponibles, y la actividad criminal en general, cuidando de no incorporar variables que pudiesen estar correlacionadas con el crimen, pero que su inclusión pudiera implicar la incorporación de sesgos no deseados y poco éticos.

3.2.6. Configuraciones previas a la predicción

3.2.6.1. Incorporación de un componente cíclico estacional

Además de las ya mencionadas variables predictoras relacionadas al número de reportes para cada categoría en períodos anteriores, y a las variables complementarias pertenecientes a cada comuna, se incluye a la semana calendario como atributo para la predicción.

La semana calendario corresponde al número identificador de la semana a la cual pertenece el reporte, con respecto al total de semanas en un año. Esta variable se incorpora como dos columnas del conjunto de datos, en su descomposición trigonométrica – con el fin de incorporar la periodicidad de esta variable, y la cercanía que tienen las semanas 52 y 1, la cual no puede ser capturada de manera lineal.

La descomposición trigonométrica que se utiliza para incluir la semana calendario como variable predictora puede ser observada en las Ecuaciones 3.2 y 3.3.

$$week\_sin = sin\left(2\pi\frac{(week-1)}{52}\right) \tag{3.2}$$

$$week\_cos = cos\left(2\pi\frac{(week-1)}{52}\right) \tag{3.3}$$

3.2.6.2. Escalamiento de los datos numéricos

Previo a utilizar el conjunto de datos modelado, dadas todas las configuraciones definidas previamente, se estandarizan y escalan los datos. Esto, con el objetivo de mantener en una escala numérica comparable a cada uno de los atributos a utilizar para la construcción del modelo predictivo.

Las estadísticas sociodemográficas por comuna se escalan mediante un $MinMaxScaler(\,)$, para cada una de las configuraciones espaciales y temporales disponibles. La implementación corresponde a aquella perteneciente al paquete scikit-learn de Python. La versión de Python utilizada es 3.8.10 y la versión de scikit-learn es 0.24.2.

Con la finalidad de darle una mayor preponderancia al modelamiento y agrupación del número de reportes por categoría, estos números no se escalan ni se estandarizan. Además, se trata de números de magnitudes más bien bajas, por lo que no sobresalen mayormente en comparación a variables cuyos valores estén entre 1 y 0. Por otra parte, la semana calendario no se escala ni se estandariza, al tratarse de una descomposición trigonométrica, y por tanto mayor o igual a cero y menor o igual a 1.

3.2.6.3. Subconjunto de entrenamiento y subconjunto de prueba

Se dividió al conjunto de datos en dos subconjuntos. Uno de estos conjuntos fue el conjunto de datos utilizado para entrenar cada uno de los modelos y algoritmos a emplear, es decir, para calibrar los parámetros necesarios, entre otras configuraciones. El segundo de los subconjuntos generados fue el conjunto de prueba, utilizado para probar la eficacia de los modelos una vez calibrados los parámetros con el conjunto de entrenamiento.

El conjunto de prueba contiene un tercio ($33,3\,\%$) de los datos del conjunto de datos original, para cada una de las configuraciones espaciales y temporales definidas.

3.2.6.4. Balanceo del conjunto de datos

El conjunto de datos se encuentra profundamente desbalanceado. Hay muchas más observaciones de semanas y grillas en las cuales no hubo reportes de la categoría asociada a actividades criminales ($y_{ij} = 0$, según lo definido previamente) que aquellas en las que sí hubo ($y_{ij} = 1$).

Dos aproximaciones se ejecutan con respecto a esta situación. La primera es eliminar del conjunto de datos todas aquellas observaciones en las cuales para aquella grilla no hubo ningún reporte, de ningún tipo, durante todos los períodos de tiempo previos considerados como predictores. Este no es un acercamiento novedoso y se ha utilizado anteriormente en la literatura de problemas del mismo tipo [18]. Esta estrategia de eliminación de observaciones es más bien agresiva, pudiendo ser por ejemplo una estrategia más inclusiva con las observaciones de datos a considerar sólo haber eliminado las observaciones que no presentasen reportes durante solamente el período previo, $i-1$. Sin embargo, mediante este acercamiento, el desbalance de los datos es más evidente, lo que puede arriesgar el desempeño del modelo predictivo a implementar.

El segundo acercamiento, ampliamente utilizado en la ciencia de datos, complementario al anterior y utilizado al mismo tiempo, es hacer *undersampling* y *oversampling*. La primera técnica se utiliza según su estrategia de *majority*, y el *oversampling* se utiliza a través de la técnica *SMOTE*.

En la Tabla B.1 de Anexos puede encontrarse una caracterización del desbalanceo del conjunto de datos luego del modelamiento empleado, diferenciando por granularidad espacial y temporal.

3.3. Implementación y evaluación de los modelos

3.3.1. Algoritmos de clasificación binaria

Se utilizan seis algoritmos de clasificación binaria, con el fin de descubrir cuál de todos posee el mejor desempeño comparativo, para cada una de las configuraciones espaciales y temporales de los datos implementada.

Los métodos de clasificación utilizados son:

1. **Máquinas de vectores de soporte**: Se utiliza en su implementación *LinearSVC* de *scikit-learn*.
2. **Árboles de Decisión**: Se utiliza la implementación *DecisionTreeClassifier* de *scikit-learn*. El criterio utilizado es el de entropía.
3. **Regresión logística**: Se utiliza la implementación de *scikit-learn*, mediante el solver *liblinear* de la función *LogisticRegression*.
4. **XGBoost**: Se utiliza, en su implementación de *scikit-learn*, el booster *gbtree*, en una evaluación de cien rondas.
5. **Naïve Bayes**: Se utiliza en su implementación *GaussianNB* de *scikit-learn*.
6. **k-vecinos Más Cercanos**: Mediante búsqueda de grilla se determina implementar con 15 vecinos la función *KNeighborsClassifier* de *scikit-learn*.

3.3.2. Métricas de desempeño

Para medir el desempeño del modelo, se calculan las cuatro métricas clásicas de problemas de clasificación binaria; *Accuracy*, *Precision*, *Recall* y F1−*Score*. También se calcula el puntaje ROC−AUC.

Todas las métricas de desempeño mencionadas se calculan para cada una de las combinaciones de configuraciones espaciales y temporales determinadas, tanto para la configuración de submuestreo de los datos (*undersampling*), como para la configuración de sobremuestreo de la clase subrepresentada (*oversampling*), para cada uno de los seis modelos presentados en la subsección anterior. Es decir, dadas esas combinaciones, se ejecutan 144 modelos diferentes basados en herramientas de aprendizaje de máquinas, para la obtención de 720 resultados en total.

3.3.3. Interpretabilidad de los resultados

Para estudiar la importancia de cada uno de los atributos, en cada una de las implementaciones de los modelos, diferenciando por granularidad temporal, se utiliza la librería $SHAP$ de Python.

Dicha librería utiliza un enfoque basado en teoría de juegos para explicar la importancia de los parámetros de los modelos [41]. En particular, se utiliza en su implementación de *TreeExplainer*, la cual se enfoca en algoritmos predictivos de clasificación.

Capítulo 4

Resultados y discusión

En el presente capítulo se muestran los resultados obtenidos según la implementación de los modelos de aprendizaje de máquinas planteados.

En la primera sección se presentan tablas de resultados para las cuatro configuraciones temporales modeladas, diferenciando por configuración espacial, método predictivo utilizado y estrategia de balanceo de datos. Las tablas se acompañan de un comentario sobre las tendencias identificadas y la evaluación de los mejores resultados para cada una de las combinaciones de configuraciones disponibles.

En la segunda sección se presentan los resultados para la relevancia de los atributos utilizados. Estos, en su mayoría, son las agrupaciones de reportes por categoría, pero también, como se presentó en el capítulo pasado, se acompañan de variables socioeconómicas disponibles, y otras. Se discute sobre tendencias observadas en los resultados, y de la validez e importancia de los mismos. Los resultados se ven presentes en cuatro figuras obtenidas mediante el método *TreeShap*.

La tercera sección presenta la discusión más amplia respecto al modelamiento de los datos que permite obtener los resultados de las primeras dos secciones. Se comenta sobre la importancia del sesgo presente en los datos en los resultados obtenidos. Se discute también sobre las diferentes opciones según las cuales podrían modelarse los datos de formas distintas, y el valor que ofrece el modelamiento escogido, tanto para la componente espacial, como para la temporal, y otras.

El capítulo finaliza con dos secciones orientadas hacia las recomendaciones para el mantenimiento de los datos en la organización, junto con otras observaciones, y con una discusión acerca de la escalabilidad de los resultados, y de los modelos obtenidos en el desarrollo del trabajo.

4.1. Resultados de los algoritmos empleados

En la presente sección se encuentran los resultados finales obtenidos, una vez implementados los modelos propuestos en el capítulo anterior.

Se presentan resultados para cada una de las cuatro configuraciones temporales propuestas; dos−semanas, semanal, tres−días y diaria, cada una para cada una de las tres granularidades espaciales. Se ejecutan los seis modelos de clasificación binaria implementados, y se evalúan mediante las cinco métricas de desempeño clásicas de modelos de clasificación binaria nombradas en el capítulo anterior.

En la Tabla 4.1 se encuentran los resultados para la configuración temporal de dos−semanas. En la tabla, las letras U y O representan la configuración de balanceo de datos (*undersampling* y *oversampling − SMOTE* respectivamente). Las siglas en la columna siguiente de la tabla representan cada uno de los seis modelos de clasificación. En la Tabla 4.2 se observan los resultados para la configuración semanal. En la Tabla 4.3 para la configuración de tres−días, y en la Tabla 4.4, para la configuración diaria.

En la Tabla B.1 de Anexos puede observarse la caracterización de los conjuntos de datos empleados; el número de filas, de columnas, y de balanceo de la clase positiva del conjunto de datos, en el caso de este problema, el porcentaje de presencia de actividades criminales en el total de observaciones. En las Tablas C.1, C.2, C.3 y C.4 de Anexos se pueden ver los tiempos de procesamiento de cada uno de los modelos incluidos en esta sección, además del tiempo de generación de cada uno de los conjuntos de datos mediante código, para las configuraciones de dos−semanas, semanal, tres−días y diaria, respectivamente.

Los mejores resultados, según la métrica de desempeño ROC−AUC, para cada una de las configuraciones temporales, se obtienen para la configuración espacial de $1,11$ $[km]$ x $0,93$ $[km]$, para el modelo *XGBoost*. Los mejores resultados, en general, se obtienen para la configuración temporal de dos−semanas, para el mismo modelo y misma configuración espacial, con un puntaje ROC−AUC de $0,8365$, una *Accuracy* de $0,8223$ y un F1−*Measure* de $0,8348$.

Tabla 4.1: Resultados finales para la temporalidad de dos−semanas

			Acc.	Prec.	Recall	F1	AUC
1ª config.	U	SVM	83,23 %	87,16 %	83,23 %	84,29 %	83,59 %
		DT	76,04 %	82,52 %	76,04 %	77,87 %	75,96 %
		LR	85,38 %	87,22 %	85,38 %	85,99 %	82,91 %
		XGB	82,23 %	87,13 %	82,23 %	83,48 %	**83,65 %**
		KNN	86,03 %	85,17 %	86,03 %	85,29 %	74,43 %
		NB	82,91 %	86,52 %	82,91 %	83,94 %	82,45 %
	O	SVM	83,73 %	87,11 %	83,73 %	84,69 %	83,39 %
		DT	81,86 %	81,83 %	81,86 %	81,85 %	72,46 %
		LR	85,27 %	87,23 %	85,27 %	85,91 %	82,99 %
		XGB	86,63 %	86,06 %	86,63 %	86,24 %	77,14 %
		KNN	86,20 %	85,44 %	86,20 %	85,59 %	75,34 %
		NB	79,85 %	86,68 %	79,85 %	81,46 %	82,85 %
2ª config.	U	SVM	79,01 %	86,44 %	79,01 %	81,21 %	78,91 %
		DT	71,99 %	82,87 %	71,99 %	75,31 %	71,24 %
		LR	81,46 %	86,09 %	81,46 %	83,00 %	77,86 %
		XGB	77,08 %	86,62 %	77,08 %	79,72 %	79,19 %
		KNN	85,09 %	84,25 %	85,09 %	84,60 %	69,68 %
		NB	78,64 %	85,57 %	78,64 %	80,80 %	77,07 %
	O	SVM	79,90 %	86,10 %	79,90 %	81,84 %	78,13 %
		DT	80,80 %	81,24 %	80,80 %	81,01 %	65,96 %
		LR	80,93 %	86,02 %	80,93 %	82,60 %	77,80 %
		XGB	86,36 %	84,68 %	86,36 %	84,81 %	66,51 %
		KNN	84,71 %	84,54 %	84,71 %	84,62 %	71,40 %
		NB	72,77 %	86,00 %	72,77 %	76,20 %	77,31 %
3ª config.	U	SVM	78,93 %	87,65 %	78,93 %	81,92 %	74,70 %
		DT	67,08 %	85,49 %	67,08 %	72,84 %	67,58 %
		LR	79,08 %	87,56 %	79,08 %	82,01 %	74,43 %
		XGB	73,78 %	88,18 %	73,78 %	78,19 %	75,82 %
		KNN	86,08 %	85,86 %	86,08 %	85,96 %	65,76 %
		NB	77,57 %	87,16 %	77,57 %	80,86 %	73,29 %
	O	SVM	79,48 %	87,37 %	79,48 %	82,26 %	73,82 %
		DT	82,84 %	83,70 %	82,84 %	83,26 %	61,36 %
		LR	78,20 %	87,59 %	78,20 %	81,39 %	74,54 %
		XGB	88,98 %	86,61 %	88,98 %	86,58 %	59,54 %
		KNN	84,95 %	86,28 %	84,95 %	85,55 %	68,36 %
		NB	67,83 %	87,42 %	67,83 %	73,51 %	72,63 %

Tabla 4.2: Resultados finales para la temporalidad semanal

			Acc.	Prec.	Recall	F1	AUC
1ª config.	U	SVM	82,58 %	86,59 %	82,58 %	83,99 %	77,36 %
		DT	71,76 %	83,93 %	71,76 %	75,45 %	71,71 %
		LR	82,04 %	87,10 %	82,04 %	83,71 %	78,81 %
		XGB	77,77 %	87,52 %	77,77 %	80,52 %	**79,95 %**
		KNN	85,48 %	85,23 %	85,48 %	85,35 %	71,00 %
		NB	78,94 %	86,94 %	78,94 %	81,35 %	78,75 %
	O	SVM	80,65 %	87,25 %	80,65 %	82,70 %	79,37 %
		DT	81,70 %	82,16 %	81,70 %	81,92 %	65,80 %
		LR	81,43 %	87,04 %	81,43 %	83,25 %	78,78 %
		XGB	87,18 %	85,62 %	87,18 %	85,89 %	67,41 %
		KNN	85,20 %	85,37 %	85,20 %	85,29 %	71,91 %
		NB	73,53 %	87,08 %	73,53 %	77,11 %	78,39 %
2ª config.	U	SVM	81,49 %	88,19 %	81,49 %	83,94 %	74,02 %
		DT	67,71 %	86,33 %	67,71 %	73,68 %	67,64 %
		LR	79,77 %	88,36 %	79,77 %	82,77 %	74,78 %
		XGB	74,21 %	88,90 %	74,21 %	78,80 %	76,08 %
		KNN	85,09 %	87,21 %	85,09 %	86,02 %	69,25 %
		NB	77,44 %	88,11 %	77,44 %	81,08 %	74,03 %
	O	SVM	78,12 %	88,45 %	78,12 %	81,62 %	75,11 %
		DT	83,89 %	84,70 %	83,89 %	84,28 %	61,28 %
		LR	78,84 %	88,39 %	78,84 %	82,12 %	74,91 %
		XGB	89,54 %	87,12 %	89,54 %	87,30 %	59,25 %
		KNN	84,07 %	87,48 %	84,07 %	85,47 %	70,79 %
		NB	68,50 %	88,30 %	68,50 %	74,38 %	73,24 %
3ª config.	U	SVM	78,70 %	90,73 %	78,70 %	83,19 %	71,49 %
		DT	64,48 %	89,54 %	64,48 %	72,90 %	64,40 %
		LR	78,15 %	90,75 %	78,15 %	82,82 %	71,58 %
		XGB	72,85 %	91,30 %	72,85 %	79,17 %	73,42 %
		KNN	86,27 %	89,86 %	86,27 %	87,83 %	66,05 %
		NB	77,45 %	90,47 %	77,45 %	82,30 %	70,21 %
	O	SVM	76,76 %	90,77 %	76,76 %	81,87 %	71,59 %
		DT	86,48 %	87,93 %	86,48 %	87,17 %	56,89 %
		LR	76,54 %	90,74 %	76,54 %	81,72 %	71,43 %
		XGB	92,56 %	89,88 %	92,56 %	89,95 %	53,57 %
		KNN	83,57 %	90,18 %	83,57 %	86,26 %	68,47 %
		NB	67,23 %	90,61 %	67,23 %	75,01 %	69,33 %

Tabla 4.3: Resultados finales para la temporalidad de tres−días

			Acc.	Prec.	Recall	F1	AUC
1ª config.	U	SVM	78,27 %	88,96 %	78,27 %	82,03 %	73,58 %
		DT	69,19 %	87,64 %	69,19 %	75,31 %	68,21 %
		LR	78,89 %	88,89 %	78,89 %	82,45 %	73,32 %
		XGB	73,58 %	89,59 %	73,58 %	78,73 %	**75,26 %**
		KNN	85,01 %	88,09 %	85,01 %	86,32 %	69,00 %
		NB	77,46 %	88,78 %	77,46 %	81,43 %	72,92 %
	O	SVM	77,57 %	88,99 %	77,57 %	81,55 %	73,67 %
		DT	84,97 %	85,65 %	84,97 %	85,30 %	59,86 %
		LR	77,81 %	88,91 %	77,81 %	81,70 %	73,40 %
		XGB	90,59 %	88,18 %	90,59 %	88,17 %	57,60 %
		KNN	82,55 %	88,36 %	82,55 %	84,82 %	70,96 %
		NB	68,65 %	88,85 %	68,65 %	74,96 %	71,90 %
2ª config.	U	SVM	77,77 %	91,64 %	77,77 %	83,01 %	70,50 %
		DT	66,07 %	90,84 %	66,07 %	74,70 %	64,73 %
		LR	77,83 %	91,68 %	77,83 %	83,06 %	70,68 %
		XGB	72,20 %	92,09 %	72,20 %	79,20 %	72,06 %
		KNN	86,65 %	90,95 %	86,65 %	88,51 %	65,67 %
		NB	77,91 %	91,39 %	77,91 %	83,07 %	69,13 %
	O	SVM	76,01 %	91,77 %	76,01 %	81,83 %	71,01 %
		DT	88,10 %	89,22 %	88,10 %	88,64 %	55,78 %
		LR	75,73 %	91,74 %	75,73 %	81,64 %	70,81 %
		XGB	93,44 %	90,42 %	93,44 %	90,89 %	51,99 %
		KNN	82,00 %	91,31 %	82,00 %	85,72 %	68,67 %
		NB	68,01 %	91,36 %	68,01 %	76,15 %	67,59 %
3ª config.	U	SVM	78,04 %	93,88 %	78,04 %	84,33 %	68,29 %
		DT	62,60 %	93,37 %	62,60 %	73,46 %	61,98 %
		LR	76,66 %	93,93 %	76,66 %	83,43 %	68,56 %
		XGB	71,66 %	94,21 %	71,66 %	80,06 %	69,91 %
		KNN	87,79 %	93,26 %	87,79 %	90,21 %	62,68 %
		NB	78,74 %	93,59 %	78,74 %	84,76 %	65,93 %
	O	SVM	74,26 %	93,93 %	74,26 %	81,84 %	68,22 %
		DT	90,74 %	92,28 %	90,74 %	91,49 %	53,92 %
		LR	74,15 %	93,92 %	74,15 %	81,76 %	68,09 %
		XGB	95,08 %	91,96 %	95,08 %	93,37 %	50,41 %
		KNN	79,09 %	93,75 %	79,09 %	84,99 %	67,26 %
		NB	70,27 %	93,46 %	70,27 %	79,09 %	63,85 %

Tabla 4.4: Resultados finales para la temporalidad diaria

			Acc.	Prec.	Recall	F1	AUC
1ª config.	U	SVM	79,40 %	94,39 %	79,40 %	85,31 %	71,53 %
		DT	64,47 %	93,86 %	64,47 %	74,94 %	64,69 %
		LR	78,23 %	94,43 %	78,23 %	84,56 %	71,70 %
		XGB	72,20 %	94,74 %	72,20 %	80,51 %	**73,16 %**
		KNN	86,72 %	93,82 %	86,72 %	89,76 %	66,45 %
		NB	82,06 %	94,01 %	82,06 %	86,95 %	68,41 %
	O	SVM	76,08 %	94,45 %	76,08 %	83,14 %	71,56 %
		DT	91,53 %	92,61 %	91,53 %	92,05 %	54,73 %
		LR	75,46 %	94,48 %	75,46 %	82,73 %	71,70 %
		XGB	95,75 %	93,21 %	95,75 %	93,78 %	50,42 %
		KNN	61,74 %	94,97 %	61,74 %	72,76 %	71,84 %
		NB	70,21 %	94,23 %	70,21 %	79,12 %	68,70 %
2ª config.	U	SVM	82,61 %	96,00 %	82,61 %	88,31 %	66,22 %
		DT	61,40 %	95,85 %	61,40 %	73,82 %	61,23 %
		LR	77,75 %	96,20 %	77,75 %	85,29 %	68,53 %
		XGB	72,35 %	96,34 %	72,35 %	81,73 %	69,44 %
		KNN	85,69 %	95,91 %	85,69 %	90,14 %	65,02 %
		NB	84,54 %	95,83 %	84,54 %	89,45 %	63,85 %
	O	SVM	72,82 %	96,22 %	72,82 %	82,05 %	67,99 %
		DT	93,71 %	95,16 %	93,71 %	94,41 %	53,14 %
		LR	73,53 %	96,21 %	73,53 %	82,53 %	67,94 %
		XGB	97,21 %	94,94 %	97,21 %	96,02 %	50,06 %
		KNN	3,23 %	97,20 %	3,23 %	1,35 %	50,29 %
		NB	74,41 %	95,90 %	74,41 %	83,12 %	63,96 %
3ª config.	U	SVM	76,64 %	97,40 %	76,64 %	85,30 %	65,98 %
		DT	58,51 %	97,21 %	58,51 %	72,33 %	59,10 %
		LR	75,78 %	97,43 %	75,78 %	84,74 %	66,36 %
		XGB	71,62 %	97,49 %	71,62 %	81,97 %	66,90 %
		KNN	83,23 %	97,27 %	83,23 %	89,38 %	63,66 %
		NB	85,93 %	97,15 %	85,93 %	90,96 %	61,10 %
	O	SVM	69,95 %	97,39 %	69,95 %	80,83 %	64,51 %
		DT	95,07 %	96,82 %	95,07 %	95,92 %	52,71 %
		LR	69,83 %	97,40 %	69,83 %	80,75 %	64,63 %
		XGB	97,36 %	96,68 %	97,36 %	97,02 %	49,95 %
		KNN	2,00 %	96,68 %	2,00 %	0,69 %	50,00 %
		NB	79,54 %	97,12 %	79,54 %	87,13 %	60,03 %

Para la evaluación de los modelos, se privilegia observar la métrica ROC−AUC, complementando con las otras métricas disponibles. Luego, se observa para qué configuración, tanto espacial como temporal, se obtiene el mejor puntaje bajo esta métrica, y luego cómo se compara el valor obtenido para las otras configuraciones, tanto espaciales como temporales. El objetivo es decidir el mejor de los seis modelos de ciencia de datos, y también las características que hacen que se obtengan estos puntajes para que resulte siendo el mejor modelo.

Bajo esta evaluación propuesta, los mejores resultados se obtienen para ***XGBoost***, y para la técnica de balanceo de datos ***undersampling***. Esta resulta ser la mejor configuración para todas las granularidades, tanto temporales como espaciales. El resultado obtenido es un resultado esperado, ya que es común que para casos similares en la literatura, *XGBoost* sea el modelo de clasificación binaria de mejor desempeño (por ejemplo, ver [42]), además de ser por sí sólo un modelo de gran desempeño y sobre todo para conjuntos de datos originalmente desbalanceados, en comparación a otros modelos clásicos de clasificación binaria.

El mayor valor para la métrica ROC−AUC se obtiene para la **configuración temporal de dos–semanas**. Esta configuración presenta los únicos puntajes para dicha métrica mayores a 0,8, todos para la primera configuración de grillas, para los modelos de SVM, LR, XGB y NB para los conjuntos con *undersampling*, y SVM, LR y NB para *oversampling*.

Curiosamente, es posible observar en los resultados de que, a pesar de que *XGBoost* tenga los mejores puntajes de desempeño de los modelos para todas las granularidades temporales, esta tendencia se difumina dependiendo de la configuración espacial de los datos, desempeñándose más pobremente a medida que aumenta la fineza de las grillas, y disminuye el tamaño de éstas. También *XGBoost* tiene un rendimiento particularmente deficiente en los conjuntos de datos modelados mediante *oversampling* en comparación a sus contrapartes modeladas con *undersampling*, o a otros modelos en las mismas configuraciones espaciales, temporales, y de balanceo de datos. Por ejemplo, en la Tabla 4.4, para la tercera configuración de grillas, todos los modelos sacan alrededor de 0,6 de ROC−AUC para *undersampling*, incluyendo *XGBoost*, y para *oversampling*, *XGBoost* es el modelo de peor desempeño, con 0,4995 en esta métrica, pese a que otros modelos mantienen la tendencia.

Puede observarse en los resultados obtenidos que, en general, los modelos tienen mejores resultados a medida que las granularidades temporales son más amplias, y que las granularidades espaciales siguen la misma tendencia. Mientras más pequeñas las grillas, y las ventanas de tiempo, peores son los resultados.

Esto puede parecer lógico, pero también contraintuitivo, dependiendo del argumento. Contraintuitivo, porque es posible afirmar que a medida que las granularidades son más pequeñas, hay muchas más casillas en los conjuntos de datos cuyo valor es cero, por abarcar menos reportes y, por tanto, los modelos podrían presentar mejores resultados artificialmente, pudiendo ser más fácil para los modelos predecir muchas casillas igual a cero, mejorando los resultados sin realmente mejorar las predicciones. Sin embargo, posiblemente lo que estén considerando los modelos al ajustar sus parámetros internamente para la predicción, es que como las granularidades más amplias, de dos–semanas, o de casillas más amplias, por ejemplo, es más posible ver más diversidad de cantidades de reportes por cada una de las distintas categorías, diversidad que permitiría a los modelos asignar mayores importancias a los distintos atribu-

tos para la predicción final de actividades criminales.

Este fenómeno abre la discusión sobre la evaluación de los modelos, porque un modelo podría resultar particularmente exitoso en el entrenamiento y evaluación de sus parámetros y resultados, pero no ser de ninguna utilidad en una aplicación real, por abarcar aspectos distintos a los necesarios para tener una verdadera utilidad práctica.

En ese mismo sentido, a nivel de implementabilidad de los resultados, se considera que el modelamiento "vencedor", de dos−semanas, no necesariamente será el más práctico. A priori, y sin haber realizado un estudio de mercado pertinente, se presupone una mayor utilidad en la implementación de los modelos modelos "intermedios" en la granularidad temporal, ya sea el modelo de granularidad temporal semanal, o el de tres−días. Cuál será más útil que otro, dependerá de los fines tácticos o estratégicos de cada posible implementación, lo cual depende a su vez del problema en el que se enmarque, y de la institución que decida implementar los modelos de predicción del crimen.

A pesar de todo, en la prevención y predicción del delito serán probablemente más beneficiosos, para la seguridad ciudadana, aquellos modelos capaces de capturar la inmediatez del acto criminal, ya que los actos delictivos fugaces son, sin duda, los actos de carácter criminal más reportados en la aplicación, en comparación a actividades criminales de mayor preparación, cuyo alcance escapa del presente estudio. Por tanto, modelos predictivos a implementar que se acerquen a la inmediatez del acto criminal lo más posible resultarán tanto más exitosos, como beneficiosos.

4.2. Relevancia de los atributos

De manera complementaria al desarrollo de los modelos se realiza un anásis de relevancia de los atributos para cada una de las implementaciones, mediante el uso de la librería $SHAP$ de Python.

En las Figuras 4.1, 4.2, 4.3 y 4.4 se encuentran los gráficos de barra obtenidos para las configuraciones temporales de dos−semanas, semanal, tres−días y diaria. Los gráficos se obtienen para la configuración espacial de grillas de $1,11\ [km]\ x\ 0,93\ [km]$, y todas según el método de balanceo de datos de *undersampling*. Se grafica, para cada una de las figuras, el método *XGBoost*, por sus destacados resultados predictivos.

En los gráficos, los atributos están listados en el eje y según orden de relevancia e importancia para los resultados del modelo. La contribución de cada uno se observa en la longitud de la barra, cuantificada en el eje de las abscisas.

4.2.1. Relevancia de atributos para la configuración de dos−semanas

Para la granularidad de dos–semanas, mediante un análisis de relevancia de los atributos según *TreeShap*, la cantidad de reportes bajo la categoría 'Actividad Sospechosa' son los más relevantes para realizar una predicción de las características que este estudio comprende. En una menor importancia relativa, se encuentran la cantidad de reportes asociados a la categoría de 'Seguridad' y 'Mascota Perdida'. También adquieren relevancia variables asociadas al presupuesto municipal por comuna.

Las categorías que más pesan en la predicción final son aquellas en períodos más cercanos, $i-14$, $i-28$ e $i-56$, todos para la categoría de 'Actividad Sospechosa'. Esto da un indicio de que existe una mayor importancia, para la predicción de un robo, en las actividades ocurridas en una grilla en semanas recién pasadas, algo que puede observarse en particular en esta configuración.

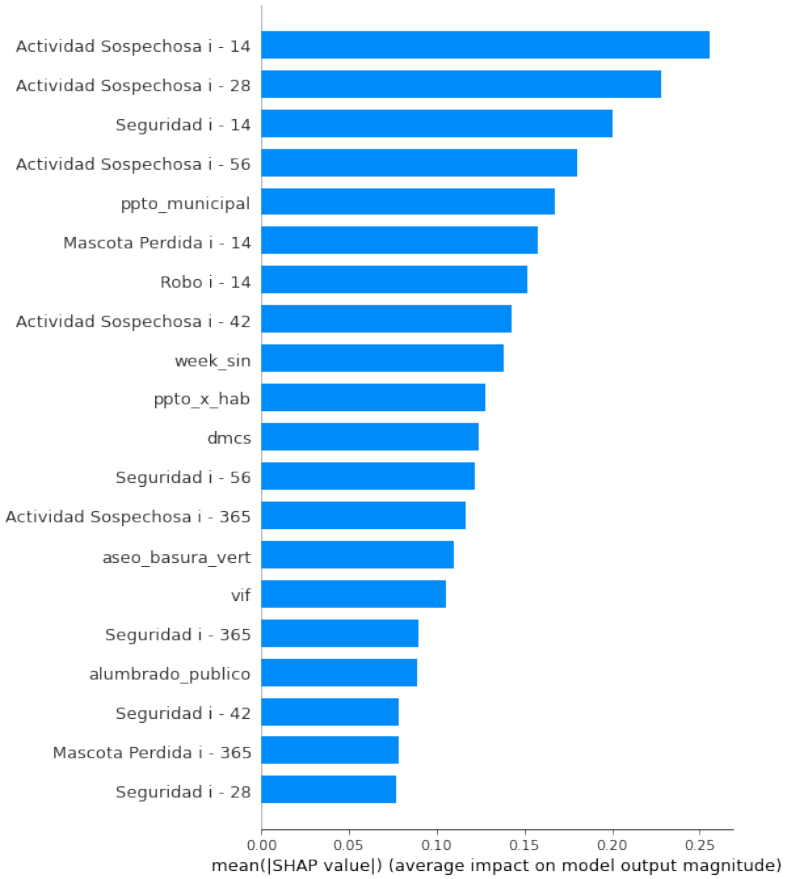

Figura 4.1: Relevancia de atributos para la configuración de dos−semanas

4.2.2. Relevancia de atributos para la configuración semanal

En la Figura 4.2 se encuentra la importancia relativa de cada uno de los atributos para la predicción en la granularidad semanal.

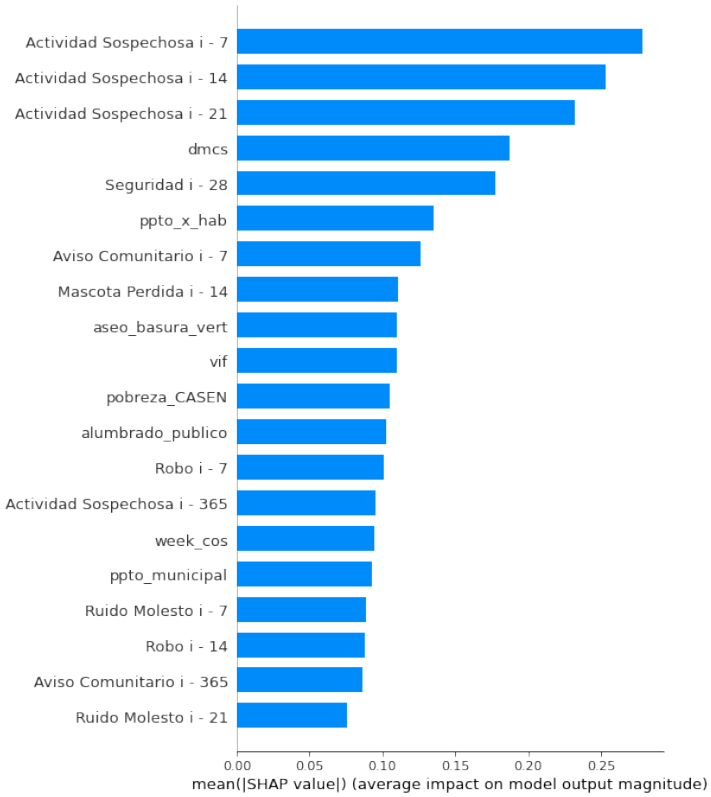

Figura 4.2: Relevancia de atributos para la configuración semanal

La mayor importancia la tienen las variables 'Actividad Sospechosa $(i-7)$', 'Actividad Sospechosa $(i-14)$', 'Actividad Sospechosa $(i-21)$', 'Aviso Comunitario $(i-7)$', presupuesto municipal por habitante, cantidad de casos policiales de Delitos de Mayor Connotación Social, y 'Seguridad' en $i-28$. Por tanto, se mantiene la tendencia de que la cantidad de reportes en 'Actividad Sospechosa', en particular en fechas recién pasadas, son más importantes comparativamente que otras, y adquieren particular importancia estadísticas sociodemográficas asociadas a la comuna.

4.2.3. Relevancia de atributos para la configuración de tres−días

En la Figura 4.3 pueden observarse los resultados para la configuración de tres–días. Esta configuración presenta el primero de los desafíos de estacionalidad, al haber incorporado variables exclusivamente para intentar capturar este efecto. Los resultados indican que se mantiene la tendencia de que 'Actividad Sospechosa', en períodos recién pasados, cobra la mayor importancia, y las estadísticas sociodemográficas por comuna, en particular, la variable correspondiente a los DMCS, adquiere cada vez más importancia relativa.

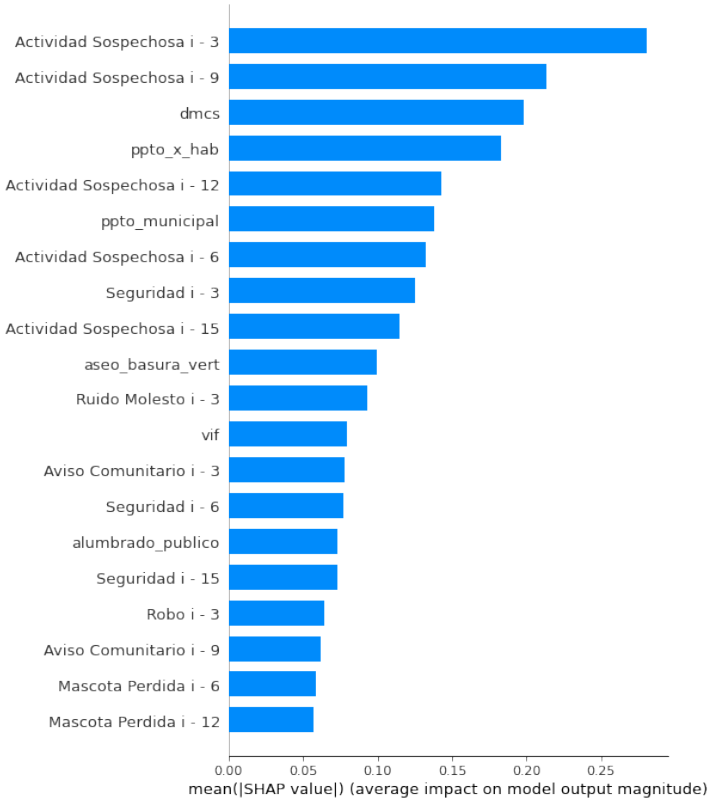

Figura 4.3: Relevancia de atributos para la configuración de tres−días

Sorprendentemente, y contrario a las hipótesis, no se observa para esta granularidad temporal la importancia de las variables exclusivamente creadas para capturar el efecto de la estacionalidad, como lo son aquellas nombradas $i-7$, $i-14$ e $i-28$. Una razón por la cual esto podría estar pasando es que el efecto de dicha estacionalidad ya esté siendo capturado por otras variables que la contienen, como aquellas asociadas a $i-6$ o $i-9$, las cuales sí tienen una importancia considerable en el gráfico presentado. También podría no existir tal efecto, pero es difícil que así sea, considerando la naturaleza del crimen en el mediano plazo.

4.2.4. Relevancia de atributos para la configuración diaria

Finalmente, en la Figura 4.4 se tienen los resultados de la importancia de cada uno de los atributos para la configuración temporal diaria. En este gráfico la variable asociada a los DMCS toma la principal relevancia, subiendo puestos gradualmente desde la primera de las configuraciones temporales estudiadas, hasta finalmente alcanzar el primer puesto. En segunda posición, la cantidad de presupuesto municipal por comuna. Luego, las variables asociadas a 'Actividad Sospechosa', para los días 1, 2, 3, 4 y 6 previos al período estudiado, no necesariamente en ese orden de relevancia.

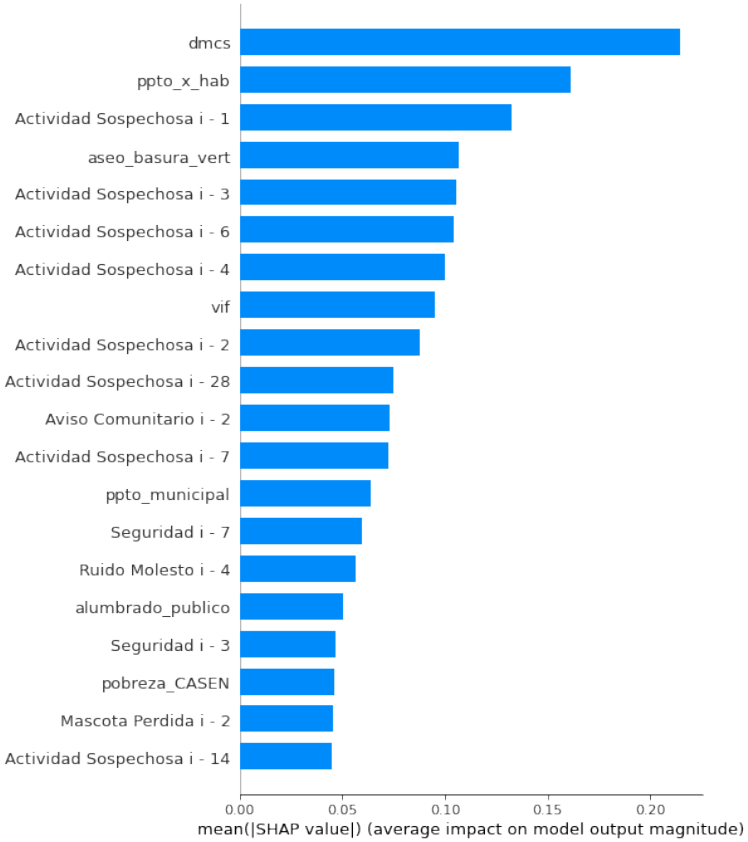

Figura 4.4: Relevancia de atributos para la configuración diaria

La principal hipótesis para que en este último caso se dé esta situación es que, al ser una configuración de mayor fineza en lo temporal, existen muchas más filas con muchos más valores iguales a cero en cada una de sus 349.396 filas, casi diez veces más que las 39.831 del caso de dos–semanas (ver Tabla B.1 de Anexos). Por tanto, es posible que los modelos les otorguen una mayor importancia a variables de la componente socioeconómica del modelo, que, comúnmente, poseen valores distintos a cero en cada una de las observaciones del conjunto de datos, por lo que los modelos encuentran algo estable en lo cual basar la estimación de sus parámetros.

El estudio de la relevancia de los atributos por granularidad temporal, y en particular las conclusiones recién propuestas, dan cuenta de la estabilidad de los modelos de granularidad temporal algo más amplia, en comparación a los más finos, por la composición que sus conjuntos de datos presentan.

4.3. Discusión

El desarrollo del presente caso de estudio y problema de investigación se lleva a cabo para eventualmente evaluar su implementación, y su utilidad, tanto para la empresa que proporciona los datos, como para entidades a un nivel superior en la jerarquía de la seguridad ciudadana, tales como lo son Carabineros, Fiscalía, o entidades privadas y públicas de todo tipo.

4.3.1. Sesgos propios de los datos

En el sentido de la implementación de los modelos que se desarrollan, hay una brecha importante de mencionar que separa su implementación de su construcción teórica y computacional. El principal supuesto a considerar es que los modelos desarrollados pueden, según un cierto nivel de precisión, predecir crímenes, delitos, robos, en general; actividades criminales. A pesar de que su función objetivo está sesgada a un fin en particular, la cual es la predicción de la ocurrencia de reportes catalogados bajo cuatro categorías diferentes, disponibles en una aplicación móvil destinada a la seguridad vecinal, se asume una correlación significativa con la predicción de actividades criminales reales, dada la asociación y pertenencia de estas categorías a la delincuencia atestiguada en la vida real. Por tanto, los modelos y resultados obtenidos pudieran sentar la base para perseguir, identificar, caracterizar, y/o prevenir actividades delictivas.

Sin embargo, inevitablemente, el modelo no escapa de su naturaleza inherente, y sólo es capaz de dibujar de cierta manera la realidad en la que se desenvuelven los delitos que el modelo abarca. El trabajo propuesto no escapa del aforismo que indica que todos los modelos se equivocan, pero hay algunos que resultan útiles [43].

En particular y en concreto, las actividades criminales que se abarcan corresponden a robos registrados en la aplicación de cuyos datos se dispone. Más aún, corresponde a robos debidamente categorizados como tales. Y, como predictores, se utilizan reportes de decenas de distintas categorías, bajo el fuerte supuesto de que aquellos se encuentran también debidamente categorizados. Esto, de hecho, puede comprobarse que no se cumple en su totalidad en el conjunto de datos, con un margen de error cuyo cálculo escapa de los alcances del problema de investigación, pero que se estima que es bajo.

Puede establecerse como hipótesis general del caso de investigación que los reportes de carácter criminal en SOSAFE, según localización, y en determinadas ventanas de tiempo, son un buen estimador de actividades delictivas observadas en el mundo real. Se trata de un estimador que posee un sesgo no nulo, y se trabaja bajo la suposición de que se trata de un sesgo bajo.

El sesgo de estimar las actividades criminales en la Región Metropolitana según las características presentadas no afecta ni el desempeño de los modelos en vista del objetivo general a resolver, ni supone un problema de escalabilidad de los modelos y parámetros determinados computacionalmente, pues, si se reducen algunos de los sesgos en una futura recopilación de datos, sean estos por ejemplo la categorización individual de los reportes, o la representatividad histórica de los datos obtenidos, la metodología empleada no diferirá en gran manera de la actualmente utilizada.

4.3.2. Modelamiento temporal y espacial de los datos

En problemas similares de predicción del crimen presentes en la literatura, el modelamiento temporal y espacial de los datos usualmente se ve determinado por las partes involucradas en el asesoramiento del proceso de modelamiento. Es decir, por ejemplo, que, si los investigadores a cargo de desarrollar un modelo de predicción del crimen están desarrollándolo por encargo de un departamento de policía, el cual los asesora de manera cercana con su experiencia y sus necesidades cotidianas, el tamaño de las grillas a modelar será determinado por las necesidades de los policías con respecto a su modelo. También es común en la literatura que sugieran o incentiven el tamaño de las ventanas temporales, puesto que estarán determinadas por las necesidades de los policías en cuanto, por ejemplo, a su asignación de recursos y sus propias restricciones institucionales. Esto ocurre de igual manera cuando quienes encargan los modelos de predicción del crimen (los *stakeholders*) son fiscalías, municipalidades, entre otros.

Por otra parte, se observa en la literatura también numerosos ejemplos y casos de uso de modelos de predicción del crimen que no han sido encargados por ninguna parte relacionada o afectada por la delincuencia, sino que son modelos que surgen como iniciativa de investigadores con el afán de crear nuevos modelos y/o arquitecturas de estado del arte, tanto de manera general como para el caso específico de predicción del crimen. El presente estudio se asemeja algo más a este último caso que al primero, en el sentido de que el objetivo no tiene encargada una implementación directa, sino que más bien se estudia la posibilidad de predecir, con un grado aceptable de precisión, la ocurrencia de actividades criminales en base a reportes de diferentes categorías, proponiendo distintos escenarios y propuestas de diseño del conjunto de datos, con la finalidad, entre otras, de comparar los resultados y la efectividad de los diseños propuestos, de los modelos, y de la premisa.

4.3.2.1. Discusiones sobre el modelamiento temporal de los datos

En casos similares de la literatura de predicción del crimen, la granularidad temporal de los modelos suele ser única, o bien unas pocas, debido a la común necesidad de una pronta implementación de los modelos desarrollados, y a la recomendación de las partes interesadas del tamaño de las ventanas temporales, justificadas según sus motivaciones particulares. En el presente caso de estudio, se presentan cuatro temporalidades distintas, entendiendo que no sirven todas para los mismos propósitos.

Bajo el supuesto de que es más probable que sea más útil predecir actividades criminales en el corto a mediano plazo, se descarta proponer ventanas temporales de predicción a largo plazo, equivalentes a uno o más años. A pesar de que hay casos en la literatura en los cuales se predice si es que habrá o no crímenes en los próximos meses [44], se escoge no perseverar en un modelamiento tan amplio.

Por otra parte, se elige no proponer granularidades temporales en el extremo corto plazo, tal como podría ser predecir la ocurrencia de crímenes en las próximas horas, en primer lugar, porque se desconfía de la eficiencia e implementabilidad de los modelos y técnicas de aprendizaje de máquinas que se pretenden emplear, en comparación a los resultados obtenidos con las granularidades algo más amplias. Además, se entiende de que los reportes, en particular aquellos delictivos, no son emitidos por parte de los usuarios justo en los momentos en que

ocurren las actividades que son precisamente reportadas, por lo que se privilegia otorgar una ventana de tiempo con una flexibilidad suficiente para capturar este fenómeno y absorber este sesgo inherente.

Por tanto, surge la propuesta formulada de cuatro temporalidades: granularidad de **dos–semanas**, granularidad **semanal**, de **tres–días**, y **diaria**. Las granularidades propuestas pueden agregarse en dos grupos; las semanales, y las diarias. Dentro de estos grupos son dos las que principalmente destacan a nivel de propuesta, una por grupo; la granularidad semanal, y la de tres–días.

La propuesta de granularidad semanal surge naturalmente en vista del objetivo de investigación; agrupar los reportes de manera tal que la suma de aquellos por categoría, más una serie de variables complementarias para la predicción, sean capaces de predecir el crimen. Frente a la necesidad de agrupar reportes por ventanas temporales, la ventana semanal surge de manera natural, por la composición impar de días en una semana, y por la gran capacidad de capturar estacionalidad, tanto a nivel anual como mensual, en la creación de variables representativas de períodos previos. Este modelamiento se orienta para capturar de manera natural las ocurrencias delictivas en el mediano plazo.

La propuesta de granularidad de tres–días, por su parte, surge por la necesidad de capturar las ocurrencias en el conjunto de datos en el corto plazo. Esta necesidad se justifica, por ejemplo, por la eventual necesidad de asignación de recursos policiales, o de distintos tipos, una vez desarrollados e implementados los modelos por alguna de las partes interesadas.

Una de las principales dificultades en el modelamiento de tres–días es que, al dividir en bloques equitativos de tres días, los períodos previos pierden gran capacidad de capturar estacionalidad en días de similares características ocurridos anteriormente. Esto se tiene puesto que se dividen las semanas (de siete días) en períodos de tres días de extensión, donde el día sobrante se ve capturado en un bloque de tres días que pertenece en su mayoría a una semana distinta. Es decir que, si se predicen actividades criminales para unos ciertos tres días de una semana, el modelo pierde la capacidad de explicar la influencia que los mismos tres días de la semana pasada tuvieron en la ocurrencia de los crímenes, teniendo sí la capacidad de decir que ocurrió los tres días previos a la predicción, y los seis, y nueve, días similares, y así sucesivamente. Este imperfecto se corrige incorporando una variable que explica lo que ocurrió en los mismos tres días de la semana previa a la de interés. Este modelamiento, sin embargo, es el único de todos los propuestos que captura información que se solapa con otros días, mas no se privilegia un modelamiento que escape de la multicolinealidad de este caso, dejando al análisis de la relevancia de los atributos del modelo la responsabilidad de indicar si fuese necesario avanzar en esta dirección.

Los otros dos modelamientos propuestos, el de dos semanas y el diario, se construyen mayormente por fines comparativos, tanto de modelamiento como de resultados, con la consideración también de que una futura implementación de dichos modelos, que aún no ha sido prevista, pudiera verse ayudada por su creación, tanto para implementaciones tácticas como estratégicas.

4.3.2.2. Discusiones sobre el modelamiento espacial de los datos

Se propone agrupar la componente espacial de los datos disponibles en grillas o celdas equitativas, de forma cuadrada.

Según la literatura, hay múltiples y diversas formas en las cuales pueden agruparse reportes policiales, según sea el caso, cuando presentan una componente espacial. Los problemas más generales, que sólo presentan ocurrencias de casos según una variable que indique distrito, comuna, provincia o región, predicen a nivel espacial la ocurrencia de delitos en dicho espacio geográfico. Aquellos estudios con disposición de datos más detallados, por ejemplo, según latitud y longitud (como es el presente caso) definen grillas equidistantes de diversas formas geométricas, o bien realizan análisis de densidad de ocurrencias del delito. Para el presente estudio se ha privilegiado el primer acercamiento con respecto al segundo, principalmente pensando en una posible implementación de la solución propuesta, y en la asignación de recursos policiales en las mencionadas grillas. El acercamiento distrital al modelamiento espacial se considera demasiado amplio para estos propósitos tácticos.

Para solucionar la ocurrencia de regiones "vacías" en la ciudad, de acuerdo con un modelamiento de tablero de ajedrez de la Región en el cual sectores verdes u de otro tipo tales como parques, plazas o estadios, entre otros, pudieran no presentar reportes de ningún tipo por la fineza de la granularidad espacial, se toma un acercamiento agresivo para evitar este sesgo en los modelos, eliminando del estudio toda celda construida en la cual no ocurran reportes de ningún tipo durante un determinado período de estudio, que puede comprender desde un mes hasta un año, dependiendo del modelamiento temporal de cada caso.

Tal como se propone en la sección 3.2.1, se modelan tres configuraciones espaciales de los datos. La primera, y original, consiste conceptualmente en construir grillas de 1 $[km]$ x 1 $[km]$, con un área de cada una de 1 $[km^2]$, que en la práctica se ve como grillas de alrededor de $1,11$ $[km]$ x $0,93$ $[km]$, por disposiciones de código y la naturaleza de los datos de latitud y longitud, ya que estas son las distancias aproximadas que se obtienen al construir grillas equiláteras de $0,01$ unidades de longitud y latitud. Por fines comparativos, y también de posible implementabilidad, se construyen dos granularidades espaciales más. La primera consiste en la mitad del área de cada grilla, y la segunda, en la mitad de alto y ancho, resultando en un cuarto de área de cada grilla. Contraintuitivamente, la creación de grillas más finas no resulta en un costo computacional aproximado más elevado, no teniendo tiempos de procesamiento significativamente superiores a la hora de construir los conjuntos de datos con el número agregado de reportes por granularidad (ver Figuras C.1 a C.4 de Anexos).

4.3.2.3. Incorporación de comunas y variables complementarias

El modelamiento y manipulación de los conjuntos de datos para entregar la posibilidad de incorporar variables dependientes de cada comuna no es trivial.

La agrupación de la cantidad de reportes por categoría, dependiendo de cada grilla o ventana temporal escogida, es una manipulación directa de los datos disponibles, agrupándolos de acuerdo a las mismas variables involucradas. La incorporación de variables socioeconómicas asociadas a cada observación, distinguiendo por comuna, implica asociar latitudes y longitudes a comunas.

En el presente conjunto de datos disponible se cuenta afortunadamente con una estandarización del sector de la ciudad en la descripción manual de la ubicación de cada reporte criminal, pero si no, es necesario construir un programa iterativo que, a nivel de recursos, puede ser muy demandante, sobre todo considerando la cantidad de comunas en el país, en la región, y dada la cantidad de datos superior a 1.700.000 reportes, para un período de sólo dos años, de los cuales el primero fue de mayormente adopción de la aplicación. En el presente caso, sólo fue necesario realizar la extracción de un string, una vez identificada la disponibilidad dentro de los datos.

La identificación de la comuna de la ubicación para cada reporte permite identificar uno de los sesgos asumidos; el de reportar en otro lugar de la ciudad. Se observan en los datos reportes emitidos desde, por ejemplo, la comuna de Lo Barnechea, denunciando mascotas perdidas en la comuna de Santiago. Incluso, reportes desde otras regiones daban cuenta de hechos ocurridos en la Región Metropolitana. Afortunadamente para los propósitos del presente caso de estudio, estas ocurrencias no superaban el 1 % de los datos, aproximadamente, por lo que no supone una dificultad de gran magnitud para fines de la predicción.

Incorporadas una vez las comunas, es posible incorporar en adelante todas las variables imaginables disponibles. Para tomar decisiones metodológicas al respecto, se espera obtener los primeros resultados, en el capítulo anterior reportados, con particular foco en la relevancia de los atributos. Sin perjuicio de aquello, se incorporan de antemano las variables que se consideran pertinentes para un primer acercamiento a la agregación de valor a los modelos, para luego, con los resultados en mano, evaluar si es pertinente agregar más, menos, o distintos predictores.

Con respecto a los atributos utilizados, todos provenientes de la base de datos del Ministerio del Interior y sus distintas subsecretarías, se cuenta con muchas más de cuatrocientas variables disponibles para incorporar a los modelos, residiendo la dificultad en la elección en la redundancia de aquellas, además de su aporte y pertinencia. Se desdeñan en una primera instancia variables tales como inversión municipal en educación, en sus distintos niveles, además de posibles redundancias de variables ya presentes, como la información de gastos municipales, gastos devengados municipales, entre otras de similar tono. Sin embargo, éstas y otras variables están disponibles para su descarga gratuita, por si fuese necesario reforzar el modelo con ellas.

Cuando una grilla comprende más de una comuna, las estadísticas comunales incorporadas a los modelos se ponderan en función de la predominancia de cada una de las comunas presentes en dicha grilla. Este enfoque metodológico permite establecer límites comunales, y una banda de flexibilidad, equivalente al diámetro de una grilla, para evitar cambios repentinos en la predicción en grillas vecinas.

4.3.3. Implementación de los modelos

Una vez incorporadas en el mismo conjunto de datos todas las variables de la componente temporal de los modelos, por categorías (cada una de las categorías para los períodos indicados en cada granularidad), se incorporan las variables socioeconómicas por comuna y la descomposición trigonométrica de la semana calendario.

Dentro de las decisiones tomadas para la implementación de los modelos, se escoge no escalar ni estandarizar la cantidad de reportes por categoría para cada uno de los períodos. Esto, con la finalidad de que adquieran preponderancia en una predicción final de los modelos, siendo este nivel de agregación el más importante del presente estudio. Para futuras implementaciones de los modelos, bien podría escoger estandarizarse, o acotarse a un cierto rango numérico que las vuelva comparables con aquellas que sí se escalaron, y quizás entreguen mejores, o peores, resultados. Las variables comunales se estandarizaron pues muchas de ellas se encontraban en la escala de los miles de millones de pesos chilenos.

Asimismo, otra de las decisiones fue la partición del conjunto de entrenamiento y del conjunto de prueba, en mantener alrededor de un tercio de los datos en el último de éstos, y los dos tercios restantes para el ajuste de los parámetros de los modelos en el primero. En futuras implementaciones, esta razón también podría modificarse, sin embargo, dada la gran escala de los datos, no se considera que se obtendrán resultados mayormente distintos a los que en este trabajo han sido presentados.

Por otra parte, la implementación de los modelos mismos fue homogénea para todas las granularidades. Hay aspectos de los resultados que pueden dar cuenta de aquello, como el particular gran desempeño de algunos modelos para algunas configuraciones, y particularmente malo para otras (ver sección 4.1 para más detalle al respecto). El foco del presente trabajo de investigación siempre estuvo en el ajuste preciso de los modelamientos espaciales y temporales, sin privilegiar demasiado el ajuste de los hiperparámetros de los modelos, mucho menos para cada caso individual, privilegiando una implementación constante y equivalente para cada uno de los casos a predecir, residiendo el mayor valor de la investigación en la correcta agregación de los reportes, y todos los modelamientos vinculados.

4.3.4. Escalabilidad de los modelos implementados

La metodología en este estudio empleada no tiene ninguna restricción por la cual no pudiera ser replicada y empleada en nuevos conjuntos de datos para distintas ciudades del país.

Es posible que los modelos, tal cual están ya ajustados y entrenados, con sus parámetros, no funcionen con grandes resultados predictivos en otras ciudades del país. Sin embargo, asociando las estadísticas sociodemográficas respectivas de cada ciudad y región, y con reportes y grillas pertenecientes a cada territorio, el estudio presenta grandes oportunidades de escalabilidad.
A nivel temporal, es posible que el modelo no presente grandes resultados, según como está entrenado el día de hoy, con nuevos datos disponibles. Un entrenamiento no presentaría mayor dificultad, ya que los tiempos de entrenamiento son, en comparación con la industria, relativamente cortos (ver Figuras C.1 a C.4 de Anexos). Es preciso recordar que los datos con

los cuales son entrenados estos modelos están históricamente sesgados, por dos eventos de singular importancia, por lo que datos de un período histórico de mayor estabilidad mejorarán su escalabilidad paramétrica. Sin embargo, la escalabilidad metodológica está asegurada, con las modificaciones que se estimen pertinentes para cada caso.

Por otra parte, la escalabilidad de los modelos frente a nuevos conjuntos de datos no está asegurada. La contribución del presente caso de estudio es su novedad con respecto a aplicaciones móviles de reportes delictivos y vecinales en general, por lo que no se asegura una escalabilidad eficiente con datos de otro tipo, por mucho que también apunten hacia la predicción de actividades criminales, como bien podrían ser datos policiales u de otro tipo. Asimismo, una implementación de los modelos en otras ciudades del planeta podría no ser directa, no sólo por la pieza clave de la investigación relativa a la agrupación de los reportes por categoría, sino que al almacenamiento y disponibilidad de datos gubernamentales por distrito para complementar la predicción en otros países. Por tanto, la implementación metodológica se incentiva, mas la implementación directa de los parámetros y modelos no se vislumbra directamente fructífera.

4.3.5. Recomendaciones para la organización

Se sugiere a la organización a cargo de la recopilación de los datos la estandarización las categorías disponibles para la elección de los usuarios de la aplicación.

Puede observarse en el análisis de los datos que muchas de las cincuenta y cinco categorías originalmente disponibles a nivel nacional son similares temáticamente entre sí, y podrían muchas de ellas contener los mismos tipos de reportes. Por ejemplo, 'Avistamiento de fauna' y 'Avistamiento de animales' podrían consolidarse dentro de la misma categoría, incluso junto con 'Incidente animal'. Hay varios casos más en este mismo sentido. Una estandarización de las categorías podría ser ventajosa para fines investigaciones y de implementación de nuevas funcionalidades de la aplicación, con un beneficio directo para los usuarios, investigadores, y últimamente, de la seguridad ciudadana en general.
Se entiende que el mencionado planteamiento posee más dimensiones que las que a este trabajo competen, por lo que corresponde a las áreas pertinentes de la empresa su consideración y eventual implementación.

El registro de la ubicación en el conjunto de datos es libremente escrito por los usuarios, otorgando datos que les parezcan relevantes como por ejemplo la dirección de la casa en la cual pudo haber sucedido un asalto. Sin embargo, existe un grado de estandarización en el string, ya que en el conjunto de datos se separa mediante una coma la ubicación escrita libremente con un sector de la ciudad estandarizado y determinado. Este sector de la ciudad, a pesar de estar estandarizado a lo largo de los datos, a nivel de ortografía, y uso de mayúsculas, por ejemplo, corresponde a divisiones político–administrativas poco formales a nivel de regulación gubernamental, por lo que se sugiere que se estandarice para un futuro empleo de los datos con fines investigativos y de desarrollo, como los que aquí se han expuesto.

Capítulo 5

Conclusiones

En el presente trabajo de investigación se han construido satisfactoriamente múltiples modelos de aprendizaje de máquinas, capaces de predecir, bajo un grado aceptable de precisión, actividades criminales, según los alcances que se han precisado y que permite el conjunto de datos, en períodos determinados, bajo un marco de implementabilidad táctica y estratégica, en la Región Metropolitana de Santiago.

Se concluye, en primer lugar, que el conjunto de datos privado, proporcionado por la organización, presenta condiciones y calidad suficiente para ser modelado y modificado, de manera tal que la información subyacente emitida por los usuarios de la aplicación móvil pueda tener utilidad para la predicción que se busca. Es posible concluir también que las herramientas de aprendizaje de máquinas propuestas cumplen exitosamente con la predicción de actividades criminales como las que se han caracterizado, logrando predecir los hechos delictivos que se reportan en la aplicación móvil, en períodos por ocurrir, en distintos sectores de la región estudiada.

Dada la novedad del origen de los datos utilizados para la construcción de los modelos, en comparación a la mayor parte de los datos disponibles para la elaboración de los modelos presentes en la literatura pertinente, se espera que los modelos propuestos sean una real contribución tanto al estado del arte, como para la real prevención del crimen, en vista de las distintas necesidades de combate a la delincuencia observadas en la actualidad, y previstas para el futuro próximo. Se concluye que el presente estudio se enmarca dentro de la literatura atingente como una contribución ética – según las precauciones y decisiones mencionadas – y relevante. Se considera que el aporte de los modelos, y la contribución del estudio presente, radica en el carácter ciudadano de las observaciones, capturando información relevante para la predicción del crimen que de otro modo pudiese ser difícil de percibir en su completitud. Además del punto de vista ciudadano, la inmediatez de la disponibilidad de los reportes, y su acceso a ellos por parte de las policías y partes interesadas, mediante la plataforma que los sostiene, permiten una contribución novedosa en este campo en particular.

De los resultados obtenidos se concluye que la granularidad temporal de mejor desempeño es aquella que agrupa en períodos de dos semanas, en comparación a las otras tres propuestas. El mejor modelo de clasificación para este caso resulta ser *XGBoost*, y la mejor técnica de balanceo de datos; undersampling. Esta configuración permite obtener una *accuracy* de $0,8223$, un F1–*Measure* de $0,8348$ y un puntaje ROC–AUC de $0,8365$.

A pesar de que la configuración de dos semanas sea la más precisa, se considera que las distintas granularidades temporales y espaciales poseen una contribución única para diversos fines, los cuales estarán supeditados a las necesidades de las partes tomadoras de decisiones, en la implementación última de los modelos. En ese sentido, las granularidades de uno y tres días resultarán adecuadas para la localización de patrullajes policiales o municipales, en diversos sectores de la región, y los modelos orientados a las predicciones semanales, se verán mayormente involucrados en la toma de decisiones a mediano plazo, para lo cual los modelos apoyarán apropiadamente la toma de decisiones.

Los modelos más precisos, según las distintas métricas de desempeño implementadas, serán aquellos que resulten más generales en cuanto a las agrupaciones de los datos, en desmedro de aquellos modelos de mayor fineza, tanto temporal como espacial, que presentarán menores precisiones en sus resultados. En general, todas las herramientas de aprendizaje de máquinas implementadas tienen buenos desempeños en sus predicciones, sin un inadecuado consumo de recursos computacionales que pudiera poner en riesgo su implementación a mayor escala. En ese mismo sentido, la escalabilidad metodológica de los modelos se considera relevante y aplicable a más casos de estudio, no así la escalabilidad paramétrica directa, la cual pudiera no ser adecuada para otros períodos, ciudades, o diferentes países.

De los resultados se concluye, además, que los reportes categorizados bajo la macrocategoría de 'Actividad Sospechosa' son los más relevantes para la predicción de actividades delictivas. Esta conclusión es altamente relevante, dado que es posible identificar una tendencia entre actos criminales y reportes emitidos por la ciudadanía en SOSAFE, lo cual no es trivial, ya que pudiera no haber tendencia, o existir una baja correlación entre una cierta variedad de reportes emitidos, o variables socioeconómicas involucradas, y los actos delictivas, lo cual no es el caso.

De los resultados se observa también que los períodos más recientes suelen tener una influencia más directa en la predicción, en desmedro de aquellos más antiguos, lo cual no supone una sorpresa que escape de una hipótesis previa. De los resultados también se obtiene que no hay evidencia suficiente para afirmar que la estacionalidad juega un rol fundamental en la predicción de actividades criminales, lo que sí escapa de hipótesis previas. Finalmente, la cantidad de casos policiales en años previos, ajustado por número de habitantes en cada comuna, resulta también de notable importancia para la predicción, lo cual puede observarse en los resultados de relevancia de los atributos. Ésta es la principal variable socioeconómica anexada observada como relevante para la predicción.

Se concluye que el presente estudio es capaz de generar modelos con una implementabilidad adecuada, satisfaciendo distintas necesidades tanto tácticas como estratégicas. Se tiene la esperanza de que contribuya a la localización de recursos, tácticos o estratégicos, policiales o de diversos tipos, para el fin último de la prevención de la delincuencia, los robos, y los actos criminales como los que en estas páginas se han descrito, a nivel regional, y eventualmente, a una escala nacional.

Capítulo 6

Trabajos Futuros

Añadir más períodos de estudio para las variables de la componente temporal

Si para la generación de los modelos predictivos se incluyeron la cantidad de reportes para cada categoría para, por ejemplo, los cuatro períodos anteriores, sería una buena dirección de avance considerar aún más períodos anteriores, tanto para capturar su relevancia como para considerar la posibilidad de mejorar los resultados predictivos de los modelos.

Incorporar más fuentes de datos complementarias

Se realizó un exhaustivo trabajo para la incorporación de estadísticas comunales, desde distintos orígenes de datos. Sería positivo evaluar la incorporación de aún más variables semejantes, pertinentes para la pregunta de investigación, y complementarias a las ya presentes.

A nivel de la componente temporal, una de las variables más directas de agregar es el reporte del clima y la temperatura diaria, por lo que esta podría ser una siguente dirección investigativa en esta misma línea. Para la componente espacial, pueden profundizarse las estadísticas comunales ya introducidas.

Evaluar la incorporación de nuevas métricas de evaluación de los modelos

Las métricas de evaluación de modelos predictivos incluidos en este trabajo son las métricas más comúnmente utilizadas al momento de trabajar con modelos de clasificación binaria. *Accuracy*, *Precision*, *Recall*, F1−*Measure* y *ROC*−*AUC* son comúnmente aceptadas como buenas métricas de desempeño. Sin embargo, existen en la literatura métricas de desempeño propias de modelos de predicción del crimen, como lo son *Hit Rate*, *Precision Accuracy Index* [45] y *Recapture Rate Index* [46]. Su incorporación como métricas de desempeño agregará una nueva dimensión a la evaluación de los modelos y profundizará su aporte dentro del contexto de la literatura de predicción del crimen, permitiendo analizar los resultados no tan sólo como un caso de uso genérico de modelos de clasificación binaria, sino que también permitiendo ponderar su capacidad predictiva a nivel espacial y temporal.

Entrenar los modelos con una mayor cantidad de datos

Los períodos de tiempo en los cuales se encuentran enmarcados los datos utilizados para calibrar los modelos no son tradicionales ni estándar. Los reportes se encuentran enmarcados primero, por las manifestaciones sociales de octubre de 2019 en Chile, y luego, por la pandemia del Covid-19, por lo que queda pendiente descubrir si sus resultados son lo suficientemente generalizables para un período de tiempo más estable dentro de la historia reciente. En cuanto haya disponibilidad de nuevos datos para fechas más recientes deberán probarse los modelos y analizar su desempeño predictivo.

Profundizar en el ajuste del desempeño de los modelos

Como se ha planteado, el foco investigativo del presente trabajo ha residido en el modelamiento y agrupación pertinente de los datos, con el fin de evaluar la utilidad para la predicción de actividades criminales. Un segundo paso en la investigación podría ser la profundización en los resultados y el desempeño en general de los modelos, una vez que se ha concluido que el acercamiento metodológico es fructífero. Modelos de mayor desempeño, arquitecturas del estado del arte, y ajuste puntual de parámetros e hiperparámetros dependiendo de cada granularidad podrían resultar en un perfeccionamiento de los actuales modelos.

Bibliografía

[1] Fundación Paz Ciudadana, "Resultados Índice Paz Ciudadana 2022," 2022, https://pazciudadana.cl/proyectos/documentos/indice-paz-ciudadana-2022/.

[2] El Mostrador, "La victimización y el miedo en Chile: disminuye la delincuencia y aumenta el temor," 2022, https://www.elmostrador.cl/noticias/opinion/2022/10/29/la-victimizacion-y-el-miedo-en-chile-disminuye-la-delincuencia-y-aumenta-el-temor/.

[3] Raphael, S. y Winter-Ebmer, R., "Identifying the effect of unemployment on crime," The journal of law and economics, vol. 44, no. 1, pp. 259–283, 2001.

[4] Basaure, M., "La delincuencia puede poner en riesgo a la democracia," 2022, https://www.ciperchile.cl/2022/10/25/la-delincuencia-puede-poner-en-riesgo-a-la-democracia/.

[5] diarioUchile, "Chile bate récord en temor al delito pese a baja victimización: ¿Quiénes son los responsables de la aparente paradoja nacional?," 2022, https://radio.uchile.cl/2022/10/27/chile-bate-record-en-temor-al-delito-pese-a-baja-victimizacion-quienes-son-los-responsables-de-la-aparente-paradoja-nacional/.

[6] Laborde, A., "Los homicidios en Chile escalan casi un 30 % en el primer semestre de 2022," 2022, https://elpais.com/chile/2022-07-13/los-homicidios-en-chile-escalan-casi-un-30-en-el-primer-semestre-de-2022.html.

[7] Samuel, A. L., "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, 1959, doi:10.1147/rd.33.0210.

[8] Fogel, D. B., Defining Artificial Intelligence. 2006.

[9] Poole, D., Mackworth, A., y Goebel, R., Computational Intelligence: A Logical Approach. New York: Oxford University Press, 1998.

[10] Turing, A. M. y Haugeland, J., "Computing machinery and intelligence," The Turing Test: Verbal Behavior as the Hallmark of Intelligence, pp. 29–56, 1950.

[11] Thompson, W., Li, H., y Bolen, A., "Artificial intelligence, machine learning, deep learning and beyond," 2022, https://www.sas.com/en_us/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html.

[12] Holland, J. H., Adaptation in Natural and Artificial Systems. University of Michigan Press, 1975.

[13] Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological review, vol. 65, no. 6, pp. 386–408, 1958, doi:https://doi.org/10.1037/h0042519.

[14] Arce, J. I. B., "La matriz de confusión y sus métricas," 2022, https://www.juanbarrios.

com/la-matriz-de-confusion-y-sus-metricas/.

[15] Salton, G. y McGill, M., Introduction to Modern Information Retrieval. International student edition, McGraw-Hill, 1983, https://books.google.cl/books?id=7f5TAAAAMAAJ.

[16] Scikit Learn, "Multiclass Receiver Operating Characteristic (ROC)," 2023, https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html.

[17] Bowyer, K. W., Chawla, N. V., Hall, L. O., y Kegelmeyer, W. P., "SMOTE: synthetic minority over-sampling technique," CoRR, vol. abs/1106.1813, 2011, http://arxiv.org/abs/1106.1813.

[18] Stalidis, P., Semertzidis, T., y Daras, P., "Examining deep learning architectures for crime classification and prediction," Forecasting, vol. 3, no. 4, pp. 741–762, 2021.

[19] Zhang, X., Liu, L., Xiao, L., y Ji, J., "Comparison of machine learning algorithms for predicting crime hotspots," IEEE Access, vol. 8, pp. 181302–181310, 2020.

[20] Kwon, E., Jung, S., y Lee, J., "Artificial neural network model development to predict theft types in consideration of environmental factors," ISPRS International Journal of Geo-Information, vol. 10, no. 2, 2021, https://www.mdpi.com/2220-9964/10/2/99.

[21] Rosenblatt, M., "Remarks on Some Nonparametric Estimates of a Density Function," The Annals of Mathematical Statistics, vol. 27, no. 3, pp. 832 – 837, 1956, doi:10.1214/aoms/1177728190.

[22] Bertsimas, D. y Kallus, N., "From predictive to prescriptive analytics," Management Science, vol. 66, no. 3, pp. 1025–1044, 2020.

[23] Thomas, A. y Sobhana, N., "A survey on crime analysis and prediction," Materials Today: Proceedings, vol. 58, pp. 310–315, 2022.

[24] Olligschlaeger, A. M., "Artificial neural networks and crime mapping," Crime mapping and crime prevention, vol. 1, p. 313, 1997.

[25] Evans, M., "New Map Room at Scotland Yard," 1947, https://www.maryevans.com/search.php?prv=preview\&job=5310922\&itm=16\&pic=11964272&row=4.

[26] Coussement, K. y Benoit, D. F., "Interpretable data science for decision making," Decision Support Systems, vol. 150, p. 113664, 2021.

[27] Peñafiel, S., Baloian, N., Sanson, H., y Pino, J. A., "Applying dempster–shafer theory for developing a flexible, accurate and interpretable classifier," Expert Systems with Applications, vol. 148, p. 113262, 2020.

[28] Baloian, N., Bassaletti, C. E., Fernández, M., Figueroa, O., Fuentes, P., Manasevich, R., Orchard, M., Peñafiel, S., Pino, J. A., y Vergara, M., "Crime prediction using patterns and context," en 2017 IEEE 21st international conference on computer supported cooperative work in design (CSCWD), pp. 2–9, IEEE, 2017.

[29] O'neil, C., Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2017.

[30] Welsh, M. y Begg, S., "What have we learned? insights from a decade of bias research," The APPEA Journal, vol. 56, no. 1, pp. 435–450, 2016.

[31] SOSAFE, "Ayudamos a construir comunidades más seguras y conectadas," 2022, es.sos

afeapp.com.

[32] Osterwalder, A. y Pigneur, Y., Business model generation: a handbook for visionaries, game changers, and challengers. John Wiley & Sons, 2010.

[33] Gob.cl, "Nuestro país," 2022, https://www.gob.cl/nuestro-pais/.

[34] Gobierno Regional Metropolitano de Santiago, "Datos geográficos," 2022, https://www.gobiernosantiago.cl/datos-geograficos/.

[35] Biblioteca del Congreso Nacional de Chile, "Región Metropolitana de Santiago," 2022, https://www.bcn.cl/siit/nuestropais/region13.

[36] Cáceres Seguel, C., "Ciudades satélites periurbanas en santiago de chile: paradojas entre la satisfacción residencial y precariedad económica del periurbanita de clase media," Revista INVI, vol. 30, no. 85, pp. 83–110, 2015.

[37] Vargas, M., "chilemapas: Mapas de las Divisiones Politicas y Administrativas de Chile (Maps of the Political and Administrative Divisions of chile)," 2022, https://pacha.dev/chilemapas/. R package version 0.3.0.

[38] Dirección del Trabajo, "Región Metropolitana de Santiago," 2022, https://www.dt.gob.cl/portal/1626/w3-article-59737.html.

[39] Centro de Estudio y Análisis del Delito, Subsecretaría de Prevención del Delito, Ministerio del Interior y Seguridad Pública, Gobierno de Chile, "Estadísticas delictuales," 2022, http://cead.spd.gov.cl/estadisticas-delictuales/.

[40] Sistema Nacional de Información Municipal, Subsecretaría de Desarrollo Regional y Administrativo, Ministerio del Interior y Seguridad Pública, Gobierno de Chile., "Datos Municipales," 2022, http://datos.sinim.gov.cl/datos_municipales.php.

[41] SHAP, "Welcome to the SHAP documentation," 2022, https://shap.readthedocs.io/en/latest/index.html.

[42] Safat, W., Asghar, S., y Gillani, S. A., "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," IEEE Access, vol. 9, pp. 70080–70094, 2021.

[43] Box, G. E., "Science and statistics," Journal of the American Statistical Association, vol. 71, no. 356, pp. 791–799, 1976.

[44] Yu, C.-H., Ward, M. W., Morabito, M., y Ding, W., "Crime forecasting using data mining techniques," 2011 IEEE 11th international conference on data mining workshops, pp. 779–786, 2011.

[45] Chainey, S., Tompson, L., y Uhlig, S., "The utility of hotspot mapping for predicting spatial patterns of crime," Security journal, vol. 21, no. 1, pp. 4–28, 2008.

[46] Levine, N., "The "Hottest" part of a hotspot: comments on "The utility of hotspot mapping for predicting spatial patterns of crime"," Security journal, vol. 21, no. 4, pp. 295–302, 2008.

[47] Biblioteca del Congreso Nacional de Chile, "Mapas vectoriales: Mapoteca," 2022, https://www.bcn.cl/siit/mapas_vectoriales/index_html.

Anexos

Anexo A. Categorías

Tabla A.1: Cantidad de reportes, según categoría, a nivel nacional

Categoría	Cantidad de reportes
Seguridad	325.035
Vandalismo	205.371
Aviso comunitario	179.445
Mascota perdida	175.192
Ruido molesto	161.765
Actividad sospechosa	139.343
Prueba	134.534
Vender	63.271
Otros públicos	53.424
Alumbrado público	40.930
Comercio habilitado	39.866
Bomberos	30.827
Buena acción	30.324
Vehículo abandonado	22.786
Accidente	22.140
Ambulancia	20.528
Basura	20.500
Comercio ambulante	16.703
Robo auto	15.921
Semáforo defectuoso	15.794
Robo a persona	13.539
Alcantarilla sin tapa	7.572
Robo casa	7.460
Comprar	7.235
Poda de Arboles	6.165
Fuga de agua	5.223
Levanta la mano	4.974
Señalética	4.382

Housekeeping	4.073
Incumplimiento del toque de queda	3.128
Estructura en mal estado	1.539
Grafitis	1.536
Situacion de calle	1.171
Sistema eléctrico en mal estado	882
Pavimento Dañado	777
Paradero en mal estado	641
Avistamiento de fauna	544
Equipamiento en mal estado	440
Residuos mal almacenados	389
Herramienta en mal estado	184
Derrame	171
EPP en mal estado	160
Avistamiento de animales	141
Animal en vía	103
Vereda en mal estado	88
Quema ilegal	83
Robo de instalaciones	78
Prueba SOSAFE	76
Pistas en mal estado	56
Gasolinera operativa	44
Adulto mayor abandonado	33
Evento climático	26
Otro	25
Alarma sonando	5
Incidente animal	1

Tabla A.2: Comunas y provincias de la Región Metropolitana

Comuna	Provincia
Colina	**Colina**
Lampa	
Tiltil	
Pirque	**Cordillera**
Puente Alto	
San José de Maipo	
Buin	**Maipo**
Calera de Tango	
Paine	
San Bernardo	
Alhué	**Melipilla**
Curacaví	
María Pinto	
Melipilla	
San Pedro	
Cerrillos	**Santiago**
Cerro Navia	
Conchalí	
El Bosque	
Estación Central	
Huechuraba	
Independencia	
La Cisterna	
La Granja	
La Florida	
La Pintana	
La Reina	
Las Condes	
Lo Barnechea	
Lo Espejo	
Lo Prado	
Macul	
Maipú	
Ñuñoa	
Pedro Aguirre Cerda	
Peñalolén	
Providencia	
Pudahuel	

Quilicura Quinta Normal Recoleta Renca San Miguel San Joaquín San Ramón Santiago Vitacura	
El Monte Isla de Maipo Padre Hurtado Peñaflor Talagante	**Talagante**

Tabla A.3: Nuevas categorías propuestas a nivel nacional

Categoría	Nueva categoría propuesta
Accidente Ambulancia Bomberos Quema ilegal	**Emergencia**
Vender Comercio habilitado Comercio ambulante Comprar	**Comercio**
Buena acción Aviso comunitario Adulto mayor abandonado	**Aviso Comunitario**
Mascota perdida Animal en vía Avistamiento de fauna Incidente animal Avistamiento de animales	**Mascota Perdida**
Robo auto Robo a persona Robo casa Robo de instalaciones	**Robo**
Prueba Prueba SOSAFE	**Prueba**
Situación de calle Vandalismo Actividad sospechosa Vehículo abandonado Grafitis	**Actividad Sospechosa**
Otros públicos Alumbrado público Basura Semáforo defectuoso Alcantarilla sin tapa Pavimento Dañado Paradero en mal estado Sistema eléctrico en mal estado Vereda en mal estado Poda de Arboles Fuga de agua	**Otros Públicos**

Señalética Levanta la mano Pistas en mal estado Gasolinera operativa Otro	
Seguridad	**Seguridad**
Ruido molesto Incumplimiento del toque de queda Alarma sonando	**Ruido Molesto**
Housekeeping Derrame Residuos mal almacenados EPP en mal estado Evento climático Equipamiento en mal estado Estructura en mal estado Herramienta en mal estado	**Minería**

Anexo B. Modelamiento del conjunto de datos

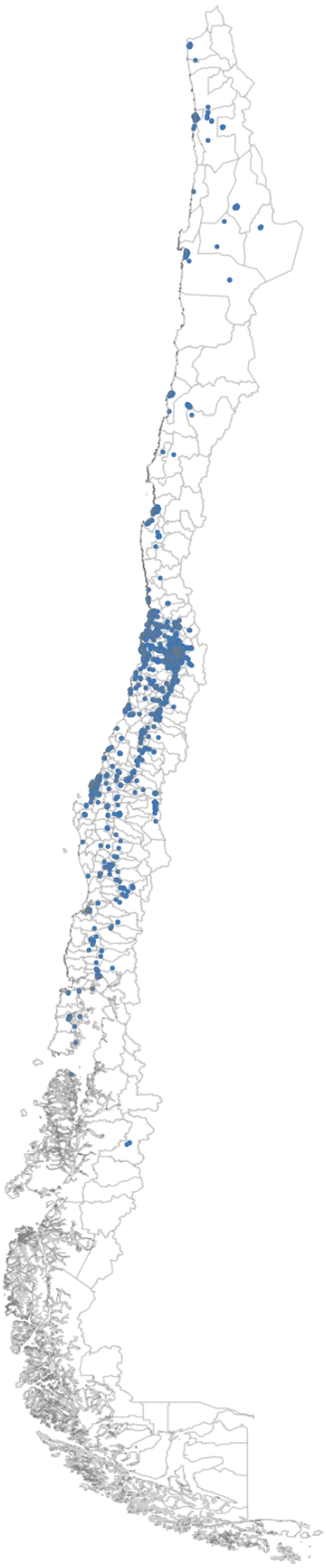

Figura B.1: Reportes a nivel nacional durante septiembre de 2019 [47]

Tabla B.1: Caracterización de conjuntos de datos modelados

$2-S$	**1ª config.**	**Filas**	39.831
		Cols.	63
		$y_{ij}=1$	13,17 %
	2ª config.	**Filas**	65.181
		Cols.	63
		$y_{ij}=1$	10,39 %
	3ª config.	**Filas**	104.969
		Cols.	63
		$y_{ij}=1$	7,75 %
S	**1ª config.**	**Filas**	69.499
		Cols.	63
		$y_{ij}=1$	8,52 %
	2ª config.	**Filas**	114.388
		Cols.	63
		$y_{ij}=1$	6,20 %
	3ª config.	**Filas**	185.202
		Cols.	63
		$y_{ij}=1$	4,37 %
$3-D$	**1ª config.**	**Filas**	142.853
		Cols.	96
		$y_{ij}=1$	4,44 %
	2ª config.	**Filas**	238.215
		Cols.	96
		$y_{ij}=1$	3,00 %
	3ª config.	**Filas**	388.289
		Cols.	96
		$y_{ij}=1$	2,01 %
D	**1ª config.**	**Filas**	349.396
		Cols.	118
		$y_{ij}=1$	1,72 %
	2ª config.	**Filas**	586.361
		Cols.	118
		$y_{ij}=1$	1,11 %
	3ª config.	**Filas**	950.375
		Cols.	118
		$y_{ij}=1$	0,72 %

Anexo C. Tiempos de procesamiento

Tabla C.1: Tiempos de procesamiento de granularidad de dos−semanas [s]

1ª config.	Generación del dataset		8,96
	U	SVM	1,55
		DT	0,25
		LR	0,93
		XGB	1,60
		KNN	0,06
		NB	2,87
	O	SVM	12,68
		DT	1,44
		LR	3,88
		XGB	8,25
		KNN	0,10
		NB	10,76
2ª config.	Generación del dataset		9,87
	U	SVM	2,52
		DT	0,36
		LR	0,40
		XGB	2,13
		KNN	0,08
		NB	6,00
	O	SVM	27,30
		DT	2,19
		LR	6,08
		XGB	14,01
		KNN	0,14
		NB	27,81
3ª config.	Generación del dataset		12,22
	U	SVM	3,34
		DT	0,51
		LR	0,99
		XGB	2,80
		KNN	0,11
		NB	11,29
	O	SVM	51,07
		DT	3,68
		LR	8,28
		XGB	22,85
		KNN	0,23
		NB	73,49

Tabla C.2: Tiempos de procesamiento de granularidad semanal [s]

1ª config.	Generación del dataset		9,63
	U	SVM	2,24
		DT	0,33
		LR	0,43
		XGB	2,16
		KNN	0,09
		NB	6,50
	O	SVM	27,81
		DT	2,44
		LR	5,81
		XGB	14,56
		KNN	0,15
		NB	31,81
2ª config.	Generación del dataset		11,12
	U	SVM	3,68
		DT	0,50
		LR	0,60
		XGB	2,79
		KNN	0,13
		NB	11,10
	O	SVM	56,77
		DT	4,57
		LR	5,63
		XGB	26,46
		KNN	0,27
		NB	105,99
3ª config.	Generación del dataset		12,94
	U	SVM	3,20
		DT	0,53
		LR	0,42
		XGB	3,52
		KNN	0,20
		NB	23,78
	O	SVM	86,60
		DT	7,85
		LR	5,41
		XGB	40,81
		KNN	0,37
		NB	227,58

Tabla C.3: Tiempos de procesamiento de granularidad de tres−días [s]

1ª config.	Generación del dataset		24,98
	U	SVM	3,03
		DT	0,46
		LR	0,56
		XGB	4,22
		KNN	0,19
		NB	13,77
	O	SVM	67,52
		DT	4,92
		LR	5,94
		XGB	38,15
		KNN	0,43
		NB	147,31
2ª config.	Generación del dataset		37,90
	U	SVM	3,24
		DT	0,53
		LR	0,52
		XGB	4,96
		KNN	0,27
		NB	32,08
	O	SVM	112,56
		DT	9,49
		LR	5,70
		XGB	62,45
		KNN	0,64
		NB	422,90
3ª config.	Generación del dataset		59,59
	U	SVM	3,11
		DT	0,66
		LR	0,41
		XGB	5,96
		KNN	0,40
		NB	47,04
	O	SVM	177,48
		DT	12,57
		LR	15,44
		XGB	102,47
		KNN	1,22
		NB	1371,33

Tabla C.4: Tiempos de procesamiento de granularidad diaria [s]

1ª config.	Generación del dataset		40,88
	U	SVM	4,44
		DT	0,82
		LR	0,62
		XGB	6,86
		KNN	0,50
		NB	42,03
	O	SVM	180,86
		DT	18,73
		LR	17,31
		XGB	124,54
		KNN	1,30
		NB	1009,98
2ª config.	Generación del dataset		60,25
	U	SVM	3,48
		DT	0,94
		LR	0,58
		XGB	8,39
		KNN	0,74
		NB	78,19
	O	SVM	305,33
		DT	29,29
		LR	10,89
		XGB	201,93
		KNN	2,26
		NB	2360,89
3ª config.	Generación del dataset		111,14
	U	SVM	2,99
		DT	1,19
		LR	0,73
		XGB	10,95
		KNN	1,21
		NB	125,59
	O	SVM	549,76
		DT	50,15
		LR	20,47
		XGB	327,48
		KNN	4,90
		NB	7495,57